ABSTRACT

Title of dissertation: MISSPECIFIED MODELS WITH PARAMETERS OF INCREASING DIMENSION

Ru Chen, Doctor of Philosophy, 2005

Dissertation directed by: Professor Eric Slud
Department of Mathematics

We study a special class of misspecified generalized linear models, where the true model is a mixed effect model but the working model is a fixed effect model with parameters of dimension increasing with sample size. We provide a sufficient condition both in linear models and generalized linear models under which the MLE derived from the misspecified working model converges to a well defined limit, and is asymptotically normal. The sample variance under the linear model is biased under model misspecification; but there exists a robust variance estimator of the MLE that converges to the true variance in probability. Criterion-based automatic model selection methods may select a linear model that contains many extra variables, but this can be avoided by using the robust variance estimator for the MLE $\hat{\boldsymbol{\beta}}_n$ in Bonferroni-adjusted model selection and by choosing λ_n that grows fast enough in Shao's GIC. Computational and simulation studies are carried out to corroborate asymptotic theoretical results as well as to calculate quantities that are not available in theoretical calculation. We find that when the link function in generalized linear

mixed models is correctly specified, the estimated parameters have entries that are close to zero except for those corresponding to the fixed effects in the true model. The estimated variance of the MLE is always smaller (in computational examples) than the true variance of the MLE, but the robust "sandwich" variance estimator can estimate the true variance very well, and extra significant variables will appear only when the link function is not correctly specified.

MISSPECIFIED MODELS WITH PARAMETERS OF INCREASNG DIMENSION

by

Ru Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

Professor Eric Slud, Chair/Advisor
Professor Francis Alt
Professor Benjemin Kedem
Professor Roland Rust
Professor Paul Smith

Dedication

Dad and Mom, this is for you.

ACKNOWLEDGMENTS

Writing this dissertation has been the most challenging task for me, but also the most fulfilling, when I look back at years of research work and the final outcome of it. I feel very lucky, for I am surrounded by so many people who inspire me, support me and more importantly, love me.

My profound gratitude toward my advisor, Dr. Slud, is from the bottom of my heart and beyond words. From the very start of this project till the very end, he has been there for me; guiding me through difficult problems, pushing me to my limit and offering me good advice in just about anything. He is the most intelligent and hard-working individual that I ever know. I am privileged to have been working with him. My thanks also go to all the professors in my dissertation committee: Dr. Francis Alt, Dr. Benjemin Kedem, Dr. Paul Smith and Dr. Roland Rust. Thank you all for being there at such an important moment, for congratulating me first, and for all the valuable suggestions to my work.

I would also like to thank my two very good friends here at the Math department in UMCP. Gaby, who has been my true friend since the first day we met and later became my inspiration; and Denise, who has always believed in me and encouraged me whenever my spirit was low. I am truly happy to have met such extraordinary girls and become their friend.

I have a wonderful crowd of friends whom I keep in touch with mainly on the

phone. They endured my complaints when I was frustrated and shared with me their own experiences down this same rocky road. I can't thank them enough for putting confidence back into my heart. In numerous dreams I went back to China to my family, whom I haven't seen for over four years now. The love and bonds among are so strong that I can still stand tall thousands of miles away, just because I know that their love is always with me no matter where I am. I also have a very loving extended family—my in-laws treated me as their own daughter from day one and have been with me all the way along. I am blessed with such friendship and love that achieving the goal makes me even more proud, because I know there are so many people that will share my joy and pride.

I saved the most important ones to the last. They are the two men I love more than anything in the world. My dad has been my idol through my pre-adult life, and continued to awe me when I grew older and understood the world better. Making him proud is the one thing I enjoyed doing over these years. I am happy that I have added something that he could be proud of for a long time. I don't know what I did in my past life to deserve a husband like Hao. He is everything a girl can possibly ask for in a man. He has seen my worst and still loves me, and he has seen my weakest and still believes in me. To have his love is to know that I have all the strength I need to pursue my dreams. I feel so lucky to have their endless love, which is the very source of all my strength and confidence. In the past few months, when I was most frustrated in my life, they taught me the true meaning of love, family and marriage. These precious memories will be with me for ever, along with my deepest love to them.

The feelings I have right now can not be described in words. I appreciate everything I have in my life. I thank the almighty hand that put me into this world and gave me such a wonderful life. I will make it worth its effort.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Background and Preliminaries

1.1 Background

"The method of maximum likelihood is, by far, the most popular technique for deriving estimators" (White[32], p.1). A fundamental assumption underlying classical asymptotic large-sample results on maximum likelihood estimation is that the stochastic law which determines the behavior of the data is known to lie within a specified parametric family of probability distributions (the proposed models). The true probability distribution (the correct model) is assumed to be one of the distributions in the specified family, or in other words the model is "correctly specified". In many situations, this might not be true. "Model misspecification" means that the specified probability family does not include the true probability law that governs the data.

Discussions of model misspecification go back to the 1960's. White [32] has a detailed account of past literature in the introductory section of his paper. He also examined the consequences and detection of model misspecification when using maximum likelihood techniques for estimation and inference. He proved under some regularity conditions in a setting with large samples of independent and identically distributed (iid) data that the estimator maximizing the working likelihood converges to a well defined limit, and gave more general robust statistics that are

analogous to Wald tests of significance for the "correctly specified" case.

The main objective of our analysis is to study a special class of misspecified models, where the true model is a mixed effect model, while the working model fails to account for the random effect, using fixed effects only. Moreover, in all but the simplest problems, some models with relatively large numbers of parameters are considered, particularly when the sample size is large. So we are allowing the dimension of the parameter space to expand at some rate less than the sample size. If we assume that different parameter values identify different probability laws in the parametric family, then the parameter space is the set including all the possible parameter values. Most statistical procedures depend heavily on asymptotic methods which rely on the central limit theorem for the parameter estimators and provide good approximations for remarkably small sample sizes when the dimension of the parameter space is fixed and not too large. When we allow the parameter space of our working model to grow with the sample size, the validity of the approximation need to be carefully examined. Portnoy [20] studied the asymptotic behavior of likelihood methods in natural exponential families when the number of parameters tends to infinity, and gave a rate at which the number of parameters can increase (compared to the sample size) so that the asymptotic distributional approximations for maximum likelihood estimators and likelihood ratio tests may be accepted as reliable. Other discussions of this sort include Strawderman and Tsiatis [27] and He and Shao [12], who focused on consistency and asymptotic normality of M-estimators when the parameter space is increasing with the sample size.

Our problem has two aspects that are in violation of assumptions of classical

statistical analysis: the misspecification of the model, and the increasing dimension of the parameter space. Apparently, the past works of White [32] and Portnoy [20] and each addressed one aspect of the problem but not the other: White [32] assumed that the data are independent and identically distributed (iid) , and the parameter space is fixed, while Portnoy[20] assumed that the data are iid and the model is correctly specified. Therefore our case is not a direct application of any of theirs. But their previous work provides a variety of tools we can use in our special situation.

With a fixed-effect working model with expanding dimension when the true model is a fixed-dimensional mixed effect model, we want to study the effects of misspecification on the number of spurious variables in model selected by automatic model selection. Various papers discussed strategies of choosing the optimal model according to certain criteria by an automatic selection procedure. The selected model will minimize (maximize) the specific criterion, and in this setting consistency and asymptotic efficiency of the final model have been well studied for models of a fixed dimension. (See, for example, Rao and Wu[21] for a list of these selection methods and their asymptotic properties.) Since any model selected will still be a fixed-effect model and thus cannot be the right one, we are more concerned about the number of spurious variables in the model, i.e. the number of variables in the selected "optimal model" that are not the true fixed effects. We want to determine if leaving out the random effect will lead us to include more variables than necessary in the final model chosen by an automatic model selection method.

1.2 Overview

The main topics we cover in this project are the effects of the specific model misspecification as outlined in Section 1.1 on parameter estimation and model selection.

Results are demonstrated in the two subsequent chapters. In Chapter 2 we focus on the normal-linear regression models, and in Chapter 3 we discuss the Generalized Linear Models with the Logistic and Poisson regression models as two special cases.

Section 2.2 discusses in detail the asymptotic behavior of the estimators derived from the working model, including the Least Squares (LS) estimator and the sample variance as a variance estimator. Results on asymptotic behavior of the estimators, when the dimension of the parameter space is fixed or when the model is correctly specified, are available from the past literature. Asymptotic analysis of the estimators under our specific setting is neither discussed elsewhere nor a direct application of past results. We use techniques of Portnoy [20] to study the conditions under which the LS estimator is still consistent and asymptotically normal. We also prove that the sample variance is a biased estimator of variance of the LS estimator, propose a robust version of the variance estimator, and prove its consistency under the model misspecification using techniques of White [32].

Sections 2.3 and 2.4 both study the effect of model misspecification on criterion-based model selection procedures. The quantity we are interested in is the expected number of extra variables in the optimal model selected by a model selection pro-

cedure. We prove in Section 2.3 that the Bonferroni-adjusted model selection procedure will choose a model that contains a number of extra variables that goes to infinity with the sample size if we use the sample variance as the variance estimator, but if we use the robust sandwich variance estimator, the experiment-wise error rate will be controlled at the right level. In Section 2.4 we study Shao's GIC, which represents a class of popular model selection methods, and conclude that the expected number of extra variables can be near zero if we let λ_n in Shao's GIC increase fast enough.

In Chapter 3, we first discuss the consistency and asymptotic normality of the MLE in a general setting in Section 3.2. We give conditions under which the MLE is consistent and asymptotically normal. These conditions are then checked in Section 3.3 and 3.4 as special cases. Unlike the normal-linear regression case where the MLE converges to the parameters in the true model, in generalized linear models the MLE converges to the point in the parameter space which minimizes the Kullback-Leibler distance between the true and the working models. We also calculate or approximate this limit in Section 3.3 and 3.4. The computation and simulation studies in Section 3.5 confirm the theoretical results and suggest results that are not theoretically available.

1.3 Notations

Throughout our analysis we assume that the data are clustered samples. Clustered samples arise frequently in practice. This clustering may be due to gathering

repeated measurements on experimental units as in longitudinal studies or may be due to subsampling the primary sampling units. The latter type of design is common in fields such as ophthalmology, where two eyes form natural clusters, and teratology, where one gathers data on all members of a litter.

The data consist of a response variable y_{ij} together with a p_n-dimensional vector of covariates $\mathbf{x}_{ij} \in \mathbf{R}^{p_n}$, that is, $\mathbf{x}_{ij}$ are row vectors of dimension p_n. The data are gathered in clusters or groups, and $i \in \{1, \cdots, m\}$ indexes clusters while $j \in \{1, \cdots n_i\}$ indexes units within clusters. Therefore, there is a one-to-one correspondence between the single and double indexing: $(i, j) \leftrightarrow t = \sum_{k=1}^{i-1} n_k + j$ for $j \in \{1, \cdots, n_i\}, \quad i \in \{1, \cdots, m\}$ and $t \in \{1, \cdots, n\}$, where n_i is the total number of objectives (responses) in the i^{th} cluster, and $n = \sum_{i=1}^{m} n_i$ is the sample size.

For the subsequent chapters, we will assume that the true model is a mixed-effect model, and the working model is a fixed-effect model. We denote by $\mathbf{X}_n^*$ the $n \times p^*$ design matrix of the true model, and $\boldsymbol{\beta}^*$ its $p^* \times 1$ fixed effects parameter vector. The random effect is assumed to be a random intercept at the cluster level. The vector of random effects of the true model is denoted by $\mathbf{u}$, a $m \times 1$ random vector. The $n \times p_n$ design matrix $\mathbf{X}_n$ of the working model includes all the columns of the true model, i.e. $\mathbf{X}_n = (\mathbf{X}_n^* | \mathbf{X}_n^0)$, where $\mathbf{X}_n^0$ is a $n \times q_n$ matrix. The $p_n \times 1$ vector $\boldsymbol{\beta}_n$ denotes the parameter vector in the working model. The number p^* of fixed effects in the true model is fixed but we will allow q_n (and therefore $p_n = p^* + q_n$) to depend on n. The situation where the design matrix of the working model does not include all the columns of the true model ("Omitted Covariates") was discussed in Neuhaus [19] and Drake and McQuarrie [8], and will not be considered here.

With clustered data, it is useful to mention two special types of covariates. The first type, a cluster-constant or cluster-level covariate, has the same value for all the units in the cluster. This is the type of covariates we assume in Chapter 2, with the additional assumption that the (distinct) cluster-level covariates come from a common distribution. The second covariate type, a designed within-cluster covariate, varies with identical distribution across the units within each cluster. This is the type of covariates we assume in Chapter 3, with further assumption that all the covariates come from a common distribution.

Since we have two types of covariates (the cluster-constant and designed within-cluster covariates), we use the notations $\tilde{\mathbf{X}}_n$ and $\tilde{n}$ instead of $\mathbf{X}_n$ and n as generic notations in this chapter. With cluster-constant covariates, $\tilde{\mathbf{X}}_n$ is the $m \times p_n$ matrix consisting of the row vectors of $\mathbf{X}_n$ from the m clusters (or, the m cluster-level covariates), and $\tilde{n} = m$. With designed within-cluster covariates, $\tilde{\mathbf{X}}_n = \mathbf{X}_n$ and $\tilde{n} = n$. The number p_n of parameters in the working model is the same in both types of covariates, so the generic notation for p_n is still p_n.

We denote by $\tilde{\mathbf{x}}_{\tilde{t}}$ for $\tilde{t} \in \{1, \cdots, \tilde{n}\}$ the $\tilde{n}$ rows of $\tilde{\mathbf{X}}_n$. Let $\mathbf{x}_{ij}$ or $\mathbf{x}_t$ be the row vectors of $\mathbf{X}_n$, let $\mathbf{x}^*_{ij}$ or $\mathbf{x}^*_t$ the row vectors of $\mathbf{X}^*_n$, for $j \in \{1, \cdots n_i\}$, $i \in \{1, \cdots, m\}$ and $t \in \{1, \cdots, n\}$. We use the pair (ij) exclusively for indices of the double-indexed responses or rows of $\mathbf{X}^*_n$ and $\mathbf{X}_n$, and t for the corresponding indices of the single-indexed responses or rows. The column vectors of $\mathbf{X}_n$ are $\mathbf{x}^{(k)}$ and the column vectors of $\tilde{\mathbf{X}}_n$ are $\tilde{\mathbf{x}}^{(k)}$ for $k \in \{1, \cdots, p_n\}$. We suppress the subscript n in these notations but they all depend on n. All boldface lowercase letters except for the letter $\mathbf{x}$ (with subscripts and/or superscripts) are column vectors. Prime of a vector

or matrix denotes transpose, while prime of a function denotes derivative.

1.4 Definitions

In this section we will list the definitions that are needed in the following chapters.

Definition 1.1 (Special Matrix Notations) *The following notations are reserved for special matrices:*

1. $\mathbf{J}_n$: *the* $n \times n$ *matrix whose elements are all* 1*'s.*

2. $\mathbf{I}_n$: *the identity matrix of size* n*.*

3. $\mathbf{1}_n$: *the* $n \times 1$ *vector whose entries are all* $1's$*.*

4. $\mathbf{D} = diag(d_1, d_2, \cdots, d_n)$: *the* $n \times n$ *diagonal matrix with* $d_1, d_2, \cdots, d_n$ *as its diagonal elements; when* $d_1, d_2, \cdots, d_n$ *are square matrices, the notation means that* $\mathbf{D}$ *is a block-diagonal matrix with* $d_1, \cdots, d_n$ *as its diagonal sub-matrices and zero matrices (of the right dimension) as its off-diagonal sub-matrices.*

Definition 1.2 (O_p, o_p, O and o) *We say:* $V_n = O_p(R_n)$ *if and only if* $V_n = R_n O_p(1)$*, where* $O_p(1)$ *denotes a sequence that is bounded in probability.* $V_n = o_p(R_n)$ *if and only if* $V_n = R_n o_p(1)$*, where* $o_p(1)$ *denotes a sequence that goes to zero in probability. Similarly, for two functions* f *and* g*,* $f = O(g)$ *means* $f(x)/g(x)$ *stays bounded as* $x \to \infty$*, and* $f = o(g)$ *means* $f(x)/g(x) \to 0$ *as* $x \to \infty$*.*

Definition 1.3 (Ordering of matrices) *We say that the $n \times n$ symmetric matrix $\mathbf{P}_1$ is less than another $n \times n$ symmetric matrix $\mathbf{P}_2$ in the matrix sense, denoted by $\mathbf{P}_1 \leq \mathbf{P}_2$, if*

$$\mathbf{v}'\mathbf{P}_1\mathbf{v} \leq \mathbf{v}'\mathbf{P}_2\mathbf{v} \tag{1.1}$$

for all unit vectors $\mathbf{v} \in \mathbf{R}_n$. The strict ordering $\mathbf{P}_1 < \mathbf{P}_2$ means that the inequalities (1.1) are strict for all unit vectors $\mathbf{v}$.

There are two immediate conclusions we can draw if $\mathbf{P}_1 \leq \mathbf{P}_2$:

1. $(\mathbf{P}_2 - \mathbf{P}_1)$ is a nonnegative definite matrix;

2. $\mathrm{tr}\mathbf{P}_1 \leq \mathrm{tr}\mathbf{P}_2$.

Definition 1.4 (L_p Norm for a random variable) *For $p > 1$, and random variable ξ, if $E|\xi|^p$ exists, then the L_p norm of ξ is $\|\xi\|_p = (E|\xi|^p)^{1/p}$.*

There are many norms defined for a vector or a matrix. In the subsequent chapters when we use the norm $\|\cdot\|$, we mean the Euclidean norm of a vector and the operator norm of a square matrix:

Definition 1.5 (Norm of A Vector) *For a vector $\mathbf{v} = (v_1, v_2, \cdots v_n) \in \mathbf{R}^n$, the Euclidean norm is*

$$\|\mathbf{v}\| = \left(\sum_{i=1}^{n} v_i^2\right)^{1/2} = \sqrt{\mathbf{v}'\mathbf{v}}.$$

Definition 1.6 (Norm of a Square Matrix) *For a square matrix $\mathbf{M}$, the operator norm is*

$$\|\mathbf{M}\| = \sup_{\|\mathbf{v}\|=1, \mathbf{v}\in\mathbf{R}^n} \mathbf{v}'\mathbf{M}\mathbf{v}.$$

For a nonnegative definite symmetric square matrix $\mathbf{M}$, all eigenvalues are nonnegative real numbers. Let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ be the largest and smallest eigenvalues of $\mathbf{M}$, or equivalently:

Definition 1.7 (Alternative Definition of $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$)

$$\begin{aligned}\lambda_{\max}(\mathbf{M}) &= \sup_{\|\mathbf{v}\|=1} \mathbf{v}'\mathbf{M}\mathbf{v} \\ \lambda_{\min}(\mathbf{M}) &= \inf_{\|\mathbf{v}\|=1} \mathbf{v}'\mathbf{M}\mathbf{v}.\end{aligned}$$

A direct application of Definition 1.7 is

$$\|\mathbf{M}\| = \lambda_{\max}(\mathbf{M}) \tag{1.2}$$

for nonnegative definite symmetric matrix $\mathbf{M}$.

Definition 1.8 ($\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}^{\otimes 2}$) *The gradient of function $f(\mathbf{v})$ with respect to vector $\mathbf{v}$ is defined by*

$$\boldsymbol{\nabla}_{\mathbf{v}} f \equiv \frac{\partial f(\mathbf{v})}{\partial \mathbf{v}}$$

and the Hessian of f with respect to $\mathbf{v}$ is defined by

$$\boldsymbol{\nabla}_{\mathbf{v}}^{\otimes 2} f \equiv \frac{\partial^2 f}{\partial \mathbf{v} \partial \mathbf{v}'}.$$

For a $n \times 1$ vector $\mathbf{v}$,

$$\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'.$$

Definition 1.9 (Modes of approximation) *For two real sequences a_n and b_n, we say*

1. *$a_n \approx b_n$ if and only if $a_n - b_n \to 0$ when $n \to \infty$.*

2. $a_n \sim b_n$ *(with* $b_n \neq 0$*) if and only if* $a_n/b_n \to 1$ *when* $n \to \infty$.

If at least one of the sequences is random, we use $\overset{p}{\approx}$ *and* $\overset{p}{\sim}$*, and the limit is in probability.*

A similar definition is available for functions:

Definition 1.10 (Modes of approximation for functions) *For two functions* $f_1(x)$ *and* $f_2(x)$ *we say*

1. $f_1(x) \approx f_2(x)$ *at* $x = x_0$ *if and only if* $\lim_{x\to x_0}(f_1(x) - f_2(x)) = 0$.

2. $f_1(x) \sim f_2(x)$ *at* $x = x_0$ *if and only if* $\lim_{x\to x_0} f_1(x)/f_2(x) = 1$.

Definition 1.11 ($\epsilon-$ Ball) *The* $\epsilon-$*ball of a vector* $\mathbf{v} \in \mathbf{R}^n$*, denoted by* $\mathbf{B}_\epsilon(\mathbf{v})$*, is the compact set*

$$\mathbf{B}_\epsilon(\mathbf{v}) = \{\mathbf{w} \in \mathbf{R}^n : \|\mathbf{w} - \mathbf{v}\| \leq \epsilon\}.$$

Definition 1.12 (Asymptotic Normality (Strong Sense)) *The* $p_n \times 1$ *estimators* $\hat{\boldsymbol{\theta}}_n$ *are said to be asymptotically normal if for any unit vectors* $\mathbf{v}_n \in \mathbf{R}^{p_n}$ *the standardized scalars* $\sqrt{n}\sigma_{\mathbf{v}_n}^{-1}\mathbf{v}_n'\hat{\boldsymbol{\theta}}_n$ *are standard Normal.*

The triangular array $w_{i,n}$ is row-independent if for each n the sequence $w_{i,n}$ for $1 \leq i \leq n$ consists of independent variables. Let $E[w_{i,n}] = 0$ for all i and n and $\sigma_n^2 = \sum_{i=1}^n E[w_{i,n}]^2$, then

Definition 1.13 (Lyapunov Condition for Triangular Arrays) *The row -independent triangular array* $w_{i,n}$ *is said to satisfy the Lyapunov condition if there exists* $\delta > 0$ *such that*

$$\frac{\sum_{k=1}^n E|w_{i,n}|^{2+\delta}}{\sigma_n^{2+\delta}} \to 0. \tag{1.3}$$

The Lyapunov condition is a sufficient condition for the Central Limit Theorem to hold(see, for example, Shiryaev [26]). For sequence $w_{t,n}$ satisfying the Lyapunov Condition (1.3), the normalized row sum converges to standard normal when $n \to \infty$.

1.5 Assumptions

Assumption 1.1 *The elements of the vector* $\boldsymbol{\beta}^* \in \mathbf{R}^{p^*}$ *satisfy* $|\boldsymbol{\beta}_i^*| > 0, \forall i \in \{1, \cdots, p^*\}$.

For large sample results, we need to control the rate at which the number of parameters in the working model is growing with the sample size, which leads to the next assumption:

Assumption 1.2 *The total number of parameters in the working model is* $p_n = p^* + q_n$, *where* $p_n = O(n^\theta)$, *with* $\theta < 1/4$. *Particularly, we assume that* $p_n = [an^\theta)]$ *where* $[x]$ *means "the greatest integer less than or equal to* x, *or* $p_n = an^\theta(1 + O(1/\log n))$ *for some constant* $a > 0$.

Our response data are not independent and identically distributed (iid), since the responses from the same cluster share the same unobservable random effect. To be able to use the large sample theory in the literature, we can view $\tilde{\mathbf{x}}_{\tilde{t}}$ as a random sample from a random vector $\boldsymbol{\xi}^{(n)}$, and discuss the problems at the cluster level, where a function of the response and the covariates is iid. Therefore we make the following assumption:

Assumption 1.3 *For each n, the rows of $\tilde{\mathbf{X}}_n$ are iid from the same distribution $F_{\boldsymbol{\xi}^{(n)}}$. Let $\boldsymbol{\xi}_k^{(n)}$ be the k^{th} element of $\boldsymbol{\xi}^{(n)}$. Then*

$$E|\boldsymbol{\xi}_k^{(n)}|^{4r} < C, \text{ for } k \in \{1, \cdots p_n\}, \text{ and } r > \frac{2\theta}{1-2\theta}.$$

That is, the $(4r)^{th}$ moments of the elements of $\boldsymbol{\xi}^{(n)}$ are uniformly bounded where r is a fixed number.

Remark: By allowing uniformly bounded higher moments on the elements of $\tilde{\mathbf{X}}_n$, the matrix $\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n$ can be better controlled when $\tilde{n}$ goes to infinity. In later discussions, we will give sufficient conditions on how large r should be for various purposes. □

Let the $p_n \times p_n$ matrix $\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)} \equiv E\left[\boldsymbol{\xi}^{(n)\otimes 2}\right]$ be defined for $1 \leq k, l \leq p_n$ by

$$\{\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\}_{kl} \equiv E[\boldsymbol{\xi}_k^{(n)}\boldsymbol{\xi}_l^{(n)}].$$

Assumption 1.4 *There exist positive constants m^* and M^* independent of n such that for every n,*

$$m^*\mathbf{I}_{p_n} \leq \boldsymbol{\Sigma}_{\mathbf{x}}^{(n)} \leq M^*\mathbf{I}_{p_n}.$$

Remark 1: Because we allow the number of parameters to go to infinity with the sample size n, we might face multicollinearity problems when there are too many parameters in the model; Assumption 1.4 bounds $\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}$ below and above so that it is always a nonsingular matrix, and as we prove later that $(\tilde{n}^{-1}\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n)$ is very close to $\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}$, the multicollinearity problem is avoided because $\tilde{\mathbf{X}}_n$ will have full rank. □

Remark 2: Assumption 1.4 also guarantees that for any $\mathbf{b} \in \mathbf{R}^{p_n}$ with $\|\mathbf{b}\| \neq 0$, the random variable $\boldsymbol{\xi}^{(n)}\mathbf{b}$ will not degenerate to 0, and for $\|\mathbf{b}\| < \infty$, the random

variable $\boldsymbol{\xi}^{(n)}\mathbf{b}$ has finite second moment. To see this, note that $E[\mathbf{b}'\boldsymbol{\xi}^{(n)'}\boldsymbol{\xi}^{(n)}\mathbf{b}] = \mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\mathbf{b}$ is bounded away from both 0 and ∞ by Assumption 1.4. □

Up to now we have not discussed the assumptions on the cluster sizes n_i. In a clustered data structure, there are two ways to increase the sample size: to increase the number of clusters ($m \to \infty$) or to increase the cluster sizes ($n_i \to \infty$). We consider the first case and view n_i as a sample from some population with finite moments. Since m goes to infinity at the same rate as n, we exchange $O(m)$ and $O(n)$ in the subsequent chapters without specific comment. Moreover, in the generic notation, $O(\tilde{n}) = O(n)$ for both types of covariates, and we do not make further comment about the difference between $O(\tilde{n})$ and $O(n)$ in this chapter, either.

Assumption 1.5 *The numbers $n_i,\ i = 1, \cdots, m$ are independent, identically distributed random variables with $1 < n_i \leq N_{\max}$ almost surely, $En_1 = N_1 > 1$ and $En_1^2 = N_2 < \infty$.*

The following assumption provides iid clusters:

Assumption 1.6 *For each n, the pairs $(\tilde{\mathbf{x}}_{\tilde{t}}, n_{\tilde{t}})$ are iid for different $\tilde{t}$.*

1.6 The Problem of Increasing Dimension

In classical statistical analysis, the number of parameters in the model is usually fixed. But in real statistical analyses the complexity of a model is often related to the size of available data. The asymptotic distribution of the parameter estimates are usually derived by taking the sample size to infinity for a fixed number

of parameters. Usually in large sample inference, when the dimension of $\mathbf{X'}_n\mathbf{X}_n$ is fixed and $\mathbf{X}_n$ has iid rows, the stability of the matrix $(n^{-1}\mathbf{X'}_n\mathbf{X}_n)$ follows from the Law of Large Numbers(LLN) applied to the p columns of $\mathbf{X}_n$ and virtually poses no additional difficulty beyond applying LLN finitely many times. When we have a parameter space of increasing dimension, this method is no longer valid. Firstly, we need to point out that when the dimension of $\mathbf{X}_n$ grows with n, the convergence of $(\tilde{n}^{-1}\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n)^{-1}$ is in the sense of operator norm. That is, $\left\|(\tilde{n}^{-1}\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n) - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right\| \to 0$ with probability approaching 1. For this "convergence in operator norm", we need the contribution (of the difference between $(n^{-1}\mathbf{X'}_n\mathbf{X}_n)$ and its expectation) from each column to be controlled to order p_n^{-1}. This can be done by assuming finite higher order moments for the entries of $\boldsymbol{\xi}^{(n)}$ and using the Burkholder Inequality (See Proposition B.1). This is similar to the approach Portnoy [20] took.

Theorem 1.1 *Under Assumptions 1.2 and 1.3, when $r > 2\theta/(1-2\theta)$, there exists a sequence $a_{\tilde{n}} \to 0$ when $\tilde{n} \to \infty$ such that with probability going to 1,*

$$\frac{1}{\tilde{n}}\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)} \leq a_{\tilde{n}}\mathbf{I}_{p_n}. \tag{1.4}$$

Proof: By the Cauchy-Schwarz inequality,

$$E|\boldsymbol{\xi}_k^{(n)}\boldsymbol{\xi}_l^{(n)}|^{2r} \leq \sqrt{E|\boldsymbol{\xi}_k^{(n)}|^{4r}E|\boldsymbol{\xi}_l^{(n)}|^{4r}} \leq C$$

for each $k, l \leq p_n$. So the sequence $\eta_{\tilde{t}}^{(kl)} = \tilde{\mathbf{x}}_{\tilde{t}k}\tilde{\mathbf{x}}_{\tilde{t}l} - \left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)_{kl}$ is iid with zero mean and finite $(2r)^{th}$ moment uniformly in k and l. By Proposition B.1 in the Appendix,

$$E\left|\frac{1}{\tilde{n}}\sum_{\tilde{t}=1}^{\tilde{n}}\eta_{\tilde{t}}^{(kl)}\right|^{2r} \leq C_{2r}^{2r}\|\eta_1^{(kl)}\|_{2r}^{2r}\tilde{n}^{-r}.$$

Let $\epsilon = r(1-2\theta) - 2\theta > 0$ and $0 < \delta < \epsilon/2r$. By Chebyshev's inequality,

$$\begin{aligned} & P\left[\left|\frac{1}{\tilde{n}}\sum_{\tilde{t}=1}^{\tilde{n}} \tilde{\mathbf{x}}_{\tilde{t}k}\tilde{\mathbf{x}}_{\tilde{t}l} - \left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)_{kl}\right| \geq \tilde{n}^{-\delta}p_n^{-1}\right] \\ = \;& P\left[\left|\frac{1}{\tilde{n}}\sum_{\tilde{t}=1}^{\tilde{n}} \eta_{\tilde{t}}^{(kl)}\right|^{2r} \geq \tilde{n}^{-2r\delta}p_n^{-2r}\right] \\ \leq \;& C_{2r}^{2r}\|\eta_1^{(kl)}\|_{2r}^{2r}\tilde{n}^{2r\delta-r}p_n^{2r}. \end{aligned} \tag{1.5}$$

Therefore,

$$\begin{aligned} & P\left[\max_{1\leq k,l\leq p_n}\left|\frac{1}{\tilde{n}}\sum_{\tilde{t}=1}^{\tilde{n}} \eta_{\tilde{t}}^{(kl)}\right| \geq \tilde{n}^{-\delta}p_n^{-1}\right] \\ \leq \;& p_n^2 P\left[\left|\frac{1}{\tilde{n}}\sum_{\tilde{t}=1}^{\tilde{n}} \eta_{\tilde{t}}^{(kl)}\right|^{2r} \geq \tilde{n}^{-2r\delta}p_n^{-2r}\right] \\ \leq \;& M_r\tilde{n}^{2\delta r-r}p_n^{2r+2} \end{aligned} \tag{1.6}$$

where M_r is a constant that depends on r but not k, l or $\tilde{n}$. Since $p_n = O(n^\theta) = O(\tilde{n}^\theta)$ and $\delta < \epsilon/2r$,

$$\tilde{n}^{2\delta r-r}p_n^{4r+2} = O(\tilde{n}^{2\delta r-r+2\theta r+2\theta}) = O(\tilde{n}^{2\delta r-\epsilon}) \to 0.$$

Let $\mathbf{M}^{(\tilde{n})} \equiv \frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}$. Then the element of $\mathbf{M}^{(\tilde{n})}$ in the k^{th} row and l^{th} column is $\mathbf{M}_{kl}^{(\tilde{n})} = \tilde{n}^{-1}\sum_{\tilde{t}=1}^{\tilde{n}}\eta_{\tilde{t}}^{(kl)}$, and so far we have proved that

$$P\left[\max_{1\leq k,l\leq p_n}\left|\mathbf{M}_{kl}^{(\tilde{n})}\right| \geq \tilde{n}^{-\delta}p_n^{-1}\right] = O(\tilde{n}^{2\delta r-\epsilon}) \to 0. \tag{1.7}$$

If $\mathbf{v} \in \mathbf{R}^{p_n}$ with $\|\mathbf{v}\| = 1$, and let v_k denote the k^{th} element of $\mathbf{v}$. By using Cauchy-Schwarz twice, we get

$$\begin{aligned} \mathbf{v}'\mathbf{M}^{(\tilde{n})}\mathbf{v} &= \sum_{k=1}^{p_n}\sum_{l=1}^{p_n}\mathbf{M}_{kl}^{(\tilde{n})}v_kv_l \\ &\leq \sqrt{\sum_{k=1}^{p_n}v_k^2\sum_{k=1}^{p_n}\left(\sum_{l=1}^{p_n}\mathbf{M}_{kl}^{(\tilde{n})}vl\right)^2} \end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\sum_{k=1}^{p_n}\left(\sum_{l=1}^{p_n}\mathbf{M}_{kl}^{(\tilde{n})}vl\right)^2} \\
&\leq \sqrt{p_n}\max_{1\leq k\leq p_n}\left|\sum_{l=1}^{p_n}\mathbf{M}_{kl}^{(\tilde{n})}v_l\right| \\
&\leq \sqrt{p_n}\max_{1\leq k\leq p_n}\sqrt{\sum_{l=1}^{p_n}v_l^2\sum_{l=1}^{p_n}\left(\mathbf{M}_{kl}^{(\tilde{n})}\right)^2} \\
&= \sqrt{p_n}\max_{1\leq k\leq p_n}\sqrt{\sum_{l=1}^{p_n}\left(\mathbf{M}_{kl}^{(\tilde{n})}\right)^2} \\
&\leq p_n\max_{1\leq k,l\leq p_n}\left|\mathbf{M}_{kl}^{(\tilde{n})}\right|, \qquad (1.8)
\end{aligned}
$$

and $\sup_{\|\mathbf{v}\|=1}\mathbf{v}'\mathbf{M}^{(\tilde{n})}\mathbf{v} \leq p_n\max_{1\leq k,l\leq p_n}\left|\mathbf{M}_{kl}^{(\tilde{n})}\right|$. Let $a_{\tilde{n}}=\tilde{n}^{-\delta}$; then

$$
\begin{aligned}
P\left[\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n-\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\geq a_{\tilde{n}}\mathbf{I}_{p_n}\right] &= P\left[\sup_{\|\mathbf{v}\|=1}\mathbf{v}'\mathbf{M}^{(\tilde{n})}\mathbf{v}\geq\tilde{n}^{-\delta}\right] \\
&\leq P\left[p_n\max_{1\leq k,l\leq p_n}\left|\mathbf{M}_{kl}^{(\tilde{n})}\right|\geq\tilde{n}^{-\delta}\right] \\
&\overset{(1.7)}{\leq} O(\tilde{n}^{2\delta r-\epsilon})\to 0 \qquad (1.9)
\end{aligned}
$$

Therefore, with probability approaching 1, there exists a sequence $a_{\tilde{n}}=\tilde{n}^{-\delta}\to 0$ such that $\mathbf{M}^{(\tilde{n})}=\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n-\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\leq a_{\tilde{n}}\mathbf{I}_{p_n}$. □

The following corollary is obvious by defining $\eta_{\tilde{t}}^{(kl)}\equiv\left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)_{kl}-\tilde{\mathbf{x}}_{tk}\tilde{\mathbf{x}}_{tl}$ and following the exact same arguments:

Corollary 1.1 *Under Assumptions 1.2 and 1.3, there exists a sequence $a_{\tilde{n}}\to 0$ when $\tilde{n}\to\infty$ such that*

$$\mathbf{\Sigma}_{\mathbf{x}}^{(n)}-\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n\leq a_{\tilde{n}}\mathbf{I}_{p_n}.$$

Theorem 1.1 and Corollary 1.1 therefore give a bound for the difference between $\tilde{n}^{-1}\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n$ and $\mathbf{\Sigma}_{\mathbf{x}}^{(n)}$:

$$-a_{\tilde{n}}\mathbf{I}_{p_n}\leq\frac{1}{\tilde{n}}\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n-\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\leq a_{\tilde{n}}\mathbf{I}_{p_n}.$$

Corollary 1.2 *Under Assumptions 1.2 and 1.3,* $\|\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}\| = O_p(a_{\tilde{n}})$.

Proof: From Theorem 1.1, with probability approaching 1,

$$\left\|\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right\| = \sup_{\|\mathbf{v}\|=1} \mathbf{v}'\left(\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)\mathbf{v} \leq \sup_{\|\mathbf{v}\|=1} \mathbf{v}' a_{\tilde{n}}\mathbf{I}_{p_n}\mathbf{v} = a_{\tilde{n}}.$$

□

Corollary 1.3 *Under Assumptions 1.2-1.4, the matrix* $\mathbf{\Sigma}_{\mathbf{x}}^{(n)}$ *is positive definite for any* n*, and for* n *large enough, the matrix* $(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n)$ *is also positive definite.*

Proof: Under Assumption 1.4, for any $\mathbf{v} \in \mathbf{R}^{p_n}$,

$$\mathbf{v}'\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\mathbf{v} \geq m^*\mathbf{v}'\mathbf{I}_{p_n}\mathbf{v} = m^* > 0,$$

so $\mathbf{\Sigma}_{\mathbf{x}}^{(n)}$is potsitive definite. Since $a_{\tilde{n}} \to 0$, for $\tilde{n}$ large enough $a_{\tilde{n}} < m^*$ and

$$\frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n \geq \mathbf{\Sigma}_{\mathbf{x}}^{(n)} - a_{\tilde{n}}\mathbf{I}_{p_n} \geq (m^* - a_{\tilde{n}})\mathbf{I}_{p_n},$$

which makes $(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n)$ potsitive definite. □

Corollary 1.4 *Under Assumptions 1.2-1.4, the eigenvalues of* $(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n)$ *are bounded below and above by* $(m^* - a_{\tilde{n}})$ *and* $(M^* + a_{\tilde{n}})$*, respectively.*

Proof: From Theorem 1.1 and proof of Corollary 1.3,

$$(m^* - a_{\tilde{n}})\mathbf{I}_{p_n} \leq \frac{1}{\tilde{n}}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n \leq (M^* + a_{\tilde{n}})\mathbf{I}_{p_n}.$$

The corollary follows then by Definitions 1.3 and 1.7. □

Corollary 1.5 *Under Assumptions 1.2-1.4,*

$$\left\|\left(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n\right)^{-1} - \left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)^{-1}\right\| = O_p(a_{\tilde{n}}).$$

Proof: By Assumption 1.4 and Definition 1.7,

$$\lambda_{\min}\left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right) \geq m^*$$

$$\left\|\left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)^{-1}\right\| = \lambda_{\min}^{-1}\left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right) \leq \frac{1}{m^*}$$

and similarly

$$\left\|\left(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n\right)^{-1}\right\| = \lambda_{\min}^{-1}\left(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n\right) \leq \frac{1}{m^* - a_{\tilde{n}}}.$$

Therefore

$$\begin{aligned}\left\|(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n)^{-1} - \left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)^{-1}\right\| &\leq \left\|\left(\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n\right)^{-1}\right\| \left\|\left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)^{-1}\right\| \left\|\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right\| \\ &\leq \frac{1}{m^*(m^* - a_{\tilde{n}})}\left\|\tilde{n}^{-1}\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n - \mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right\| = O_p(a_{\tilde{n}}). \quad (1.10)\end{aligned}$$

□

In later chapters, when we discuss the asymptotic limit of $p_n \times p_n$ matrices, we often need the quantity $\max_{t\leq n, k\leq p_n}|\tilde{\mathbf{x}}_{tk}|$ to be bounded in probability by some power of $\tilde{n}$. We need the following lemma on th maximum of independent variables to prove Theorem 1.2.

Lemma 1.1 *Let $w_{t,n}$ be a row-independent triangular array, where for each n, the sequence $w_{t,n}$ is iid for $1 \leq t \leq n$. If there exists a constant $C > 0$ such that $E|w_{t,n}|^p \leq C < \infty$, then for any $\delta_w > 0$,*

$$\max_{1\leq t\leq n}|w_{t,n}| = O_p(n^{1/p+\delta_w}).$$

Proof For any $\delta_w > 0$, $\varepsilon \equiv 1/p + \delta_w > 1/p$ and constant $K > 0$, we have

$$P[\max_{1\leq t\leq n}|w_{t,n}| > Kn^{1/p+\delta_w}] = 1 - P[\max_{1\leq t\leq n}|w_{t,n}| \leq Kn^{\varepsilon}]$$

$$
\begin{aligned}
&= 1 - (P[|w_{1,n}| \leq Kn^{\varepsilon}])^n \\
&= 1 - (1 - P[|w_{1,n}|^p > K^p n^{\varepsilon p}])^n \\
&\leq 1 - \left(1 - \frac{E|w_{1,n}|^p}{K^p n^{\varepsilon p}}\right)^n \\
&= 1 - \left(1 - \frac{nE[w_{1,n}|^p}{K^p n^{\varepsilon p}} + O(n^{2-2\varepsilon p})\right) \\
&= O(n^{1-\varepsilon p}) \to 0
\end{aligned}
$$

for $\varepsilon > 1/p$. □

Theorem 1.2 *For any* $\delta_1 > 0$,

$$\max_{\tilde{t}} \|\tilde{\mathbf{x}}_{\tilde{t}}\|^2 = O_p(n^{\frac{1}{2r}+\delta_1} p_n).$$

Proof: Let $w_{\tilde{t},\tilde{n}} \equiv \|\tilde{\mathbf{x}}_{\tilde{t}}\|^2/p_n$. Then $w_{\tilde{t},\tilde{n}}$ is a row-independent triangular array.

$$
\begin{aligned}
E|w_{\tilde{t},\tilde{n}}|^{2r} &= p_n^{-2r} \left(E\|\tilde{\mathbf{x}}_{\tilde{t}}\|^2\right)^{2r} \\
&= p_n^{-2r} \left(\sum_{k=1}^{p_n} E\tilde{\mathbf{x}}_{\tilde{t}k}^2\right)^{2r} \\
&\leq p_n^{-2r} p_n^{2r} \max_k (E\tilde{\mathbf{x}}_{\tilde{t}k}^2)^{2r} \\
&\leq \max_k E|\tilde{\mathbf{x}}_{\tilde{t}k}|^{4r} \leq C;
\end{aligned}
$$

So according to Lemma 1.1, for any $\delta_1 > 1$, $\max_{\tilde{t}} |w_{\tilde{t},\tilde{n}}| = O_p(n^{\frac{1}{2r}+\delta_1})$, or

$$\max_{\tilde{t}} \|\tilde{\mathbf{x}}_{\tilde{t}}\|^2 = O_p(n^{\frac{1}{2r}+\delta_1} p_n).$$

□

Chapter 2

Linear Models

Linear models are a special type of models that have an appealingly simple and interpretable analysis. This is the most widely treated branch of statistics, both in theory and in practice. The response variables are ordinarily assumed to be linear combinations of the regressors of fixed dimension plus iid zero-mean normal errors. We are interested in violating these assumptions in three ways: the responses are not independent under the true model; the number of regressors are not fixed; and the model is not correctly specified. We want to show the effect of omitting a random intercept in a normal linear model on parameter estimation, hypothesis testing and model selection.

2.1 General Notations and Assumptions

The True Model: We assume that the true model is a mixed-effect linear model with a random intercept:

$$\mathbf{y}_n = \mathbf{X}_n^* \boldsymbol{\beta}^* + \mathbf{Z}_n \mathbf{u} + \mathbf{e}, \quad \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_u^2 \mathbf{I}_m & \mathbf{0}_{m \times n} \\ \mathbf{0}_{n \times m} & \sigma_e^2 \mathbf{I}_n \end{pmatrix} \right), \tag{2.1}$$

where

$$\mathbf{Z}_n = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ & & \cdots & \\ \mathbf{0} & & \cdots & \mathbf{1}_{n_m} \end{pmatrix}_{n \times m}, \tag{2.2}$$

$\mathbf{y}_n$ is the $n \times 1$ vector of responses, $\mathbf{X}_n^*$ is the $n \times p^*$ design matrix, and the $p^* \times 1$ vector $\boldsymbol{\beta}^*$ is the coefficients of the fixed effects and $\mathbf{u}$ is a $m \times 1$ vector of iid normal random variables with mean 0 and variance σ_u^2.

The Working Model: We assume that the working model is a standard fixed-effect linear model:

$$\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta}_n + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2 \mathbf{I}_n\right) \tag{2.3}$$

where the $n \times p_n$ matrix $\mathbf{X}_n$ is the design matrix of the working model, and the $p_n \times 1$ vector $\boldsymbol{\beta}_n$ is the vector of coefficients of the fixed effects. Also the working model assumes that the entries of the $n \times 1$ vector $\boldsymbol{\epsilon}$ are iid normal variables.

As we mentioned in Chapter 1, we assume in the linear models that we have cluster-level covariates. This means that

Assumption 2.1 *With $\mathbf{Z}_n$ defined in (2.2),*

$$\mathbf{X}_n = \mathbf{Z}_n \tilde{\mathbf{X}}_n. \tag{2.4}$$

Under Assumption 2.1, we have $\mathbf{x}_{ij} = \tilde{\mathbf{x}}_i$ for $1 \leq j \leq n_i$. According to Assumption 1.6, the pair $(\tilde{\mathbf{x}}_i, n_i)$, $1 \leq i \leq m$ are iid $1 \times (p_n + 1)$ row vectors, with

$\tilde{\mathbf{x}}_i$ satisfying Assumption 1.3 and n_i satisfying Assumption 1.5. Therefore, we can define two matrices:

$$\left\{\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right\}_{kl} \equiv E[n_i \boldsymbol{\xi}_k^{(n)} \boldsymbol{\xi}_l^{(n)}]$$

and

$$\left\{\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)}\right\}_{kl} \equiv E[n_1^2 \boldsymbol{\xi}_k^{(n)} \boldsymbol{\xi}_l^{(n)}].$$

Throughout this chapter we discuss only the Least Square (LS) estimator of $\boldsymbol{\beta}_n$ (which, in normal linear regression, is also the Maximum Likelihood Estimator under the working model):

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' \mathbf{y}_n.$$

2.2 Asymptotic Behavior of the Estimators

Asymptotic behavior of the LS estimators under regularity conditions are well studied in the literature. These conditions include the independence of the data, a fixed dimension and normal errors. Various papers try to relax these assumptions. Lai, Robbins and Wei [15] discussed the strong consistency of least squares estimates in multiple regression with independent errors under minimal assumptions on the design and weak moment conditions on the errors, and Eicker [9] relaxed the identically distributed assumption on the errors. But neither discussed the problem with increasing dimension. We will first establish the stability of $(n^{-1}\mathbf{X}_n'\mathbf{X}_n)$ under Assumptions 1.2-1.6, and then prove the consistency of the LS estimators.

Theorem 2.1 *Under Assumptions 1.2-1.6, with probability going to 1,*

$$-m^{-\delta}\mathbf{I}_{p_n} \leq m^{-1}\mathbf{X}_n'\mathbf{X}_n - \boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)} \leq m^{-\delta}\mathbf{I}_{p_n},$$

and

$$-m^{-\delta}\mathbf{I}_{p_n} \leq m^{-1}\mathbf{X}_n'\mathbf{Z}_n\mathbf{Z}_n'\mathbf{X}_n - \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} \leq m^{-\delta}\mathbf{I}_{p_n},$$

where $0 < \delta < \epsilon/2r$ *is arbitrary and* $\epsilon = r(1-2\theta) - 2\theta > 0$.

Proof: For $\mathbf{Z}_n$ defined in (2.2), the following are true:

$$\mathbf{Z}_n'\mathbf{Z}_n = \text{diag}(n_1, \cdots, n_m) \equiv \begin{pmatrix} n_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & n_m \end{pmatrix}_{m\times m}, \tag{2.5}$$

and

$$\mathbf{Z}_n\mathbf{Z}_n' = \text{diag}(\mathbf{J}_{n_1}, \cdots, \mathbf{J}_{n_m}) \equiv \begin{pmatrix} \mathbf{J}_{n_1} & \cdots & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & \cdots & \mathbf{J}_{n_m} \end{pmatrix}_{n\times n} \tag{2.6}$$

where $\mathbf{J}_{n_i}$ is the $n_i \times n_i$ matrix whose elements are all 1's. With the repeated rows,

$$\mathbf{X}_n'\mathbf{X}_n = \tilde{\mathbf{X}}_n'\mathbf{Z}_n'\mathbf{Z}_n\tilde{\mathbf{X}}_n = \sum_{i=1}^{m} n_i\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i,$$

and

$$\mathbf{X}_n'\mathbf{Z}_n\mathbf{Z}_n'\mathbf{X}_n = \sum_{i=1}^{m} n_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i.$$

Since the pair $(\mathbf{x}_i, n_i)$ are iid, we can use the same arguments as in the proof of Theorem 1.1 with the new iid random variables $\sqrt{n_i}\tilde{\mathbf{x}}_i$ and $n_i\tilde{\mathbf{x}}_i$ instead of $\tilde{\mathbf{x}}_{\tilde{i}}$. The only thing we have to prove, is that the $(4r)^{th}$ moments of $\sqrt{n_i}\tilde{\mathbf{x}}_{ik}$ and $n_i\tilde{\mathbf{x}}_{ik}$ are uniformly bounded for all i and k. Under Assumption 1.5, $n_i \leq N_{\max}$ almost surely; with Assumption 1.3, $E|\sqrt{n_i}\mathbf{x}_{ik}|^{4r} \leq N_{\max}^{2r}E|\tilde{\mathbf{x}}_{ik}|^{4r} \leq N_{\max}^{2r}C$, and $E|n_i\tilde{\mathbf{x}}_{ik}|^{4r} \leq N_{\max}^{4r}C$. □

Remark: With Theorem 2.1, we have proved the stability of the two matrices $(m^{-1}\mathbf{X}'_n\mathbf{X}_n)$ and $(m^{-1}\mathbf{X}'_n\mathbf{Z}_n\mathbf{Z}'_n\mathbf{X}_n)$. Moreover, we can bound the two expectation matrices $\mathbf{\Sigma}^{(n)}_{\mathbf{x},1}$ and $\mathbf{\Sigma}^{(n)}_{\mathbf{x},2}$ in terms of bounds for $\mathbf{\Sigma}^{(n)}_{\mathbf{x}}$ and n_i :

$$m^*\mathbf{I}_{p_n} \leq \mathbf{\Sigma}^{(n)}_{\mathbf{x},1} \leq M^*N_{\max}\mathbf{I}_{p_n} \tag{2.7}$$

and

$$m^*N_{\max}\mathbf{I}_{p_n} \leq \mathbf{\Sigma}^{(n)}_{\mathbf{x},2} \leq M^*N^2_{\max}\mathbf{I}_{p_n}. \tag{2.8}$$

Then similar to Corollary 1.1 to 1.5, we can find the bounds (in the sense of ordering in matrices) for $(m^{-1}\mathbf{X}'_n\mathbf{X}_n)$ and $(m^{-1}\mathbf{X}'_n\mathbf{Z}_n\mathbf{Z}'_n\mathbf{X}_n)$. □

2.2.1 Consistency of the LS Estimator

We show in this section that in the model misspecification of (2.1) and (2.3), the least squares estimator for the coefficients $\hat{\boldsymbol{\beta}}_n$ is still consistent. Let

$$\boldsymbol{\beta}_0 = \begin{pmatrix} \boldsymbol{\beta}^* \\ \mathbf{0} \end{pmatrix}_{p_n \times 1},$$

then

$$\mathbf{X}^*_n\boldsymbol{\beta}^* = \mathbf{X}_n\boldsymbol{\beta}_0. \tag{2.9}$$

The following theorem provides a version of consistency for $\hat{\boldsymbol{\beta}}_n$. For simplicity, we drop the subscript n in $\mathbf{X}_n$, y_n and $\mathbf{Z}_n$, but bear in mind that they all depend on n.

Theorem 2.2 *Let $\hat{\boldsymbol{\beta}}_n$ be the Least Squares Estimator of the working model. Then under Assumptions 1.2-1.6 and 2.1, $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(p_n/\sqrt{n})$ as $n \to \infty$, where the probability is taken under the true model.*

Proof: First of all,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Z}\mathbf{u} + \mathbf{e}) \\ &= \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Z}\mathbf{u} + \mathbf{e}),\end{aligned}$$

so

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 = (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = (\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'(\mathbf{Z}\mathbf{u} + \mathbf{e}) \equiv v_n > 0,$$

and for constant $K > 0$,

$$P\left[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \frac{Kp_n}{\sqrt{n}}\right] = P\left[v_n > \frac{K^2p_n^2}{n}\right] \leq \frac{Ev_nn}{K^2p_n^2}.$$

To have $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(p_n/\sqrt{n})$, it suffices to prove that $Ev_nn/p_n^2 \to 0$ as $n \to \infty$.

Since $[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}']$ is nonnegative definite and the diagonal matrix $\mathbf{Z}'\mathbf{Z} \leq N_{\max}\mathbf{I}_m$, Corollary A.1 implies

$$\mathrm{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\right] \leq N_{\max}\mathrm{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'\right] = N_{\max}\mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\right].$$

The random vector $(\mathbf{Z}\mathbf{u} + \mathbf{e})$ has variance-covariance matrix

$$\mathrm{var}(\mathbf{Z}\mathbf{u} + \mathbf{e}) = \sigma_u^2\mathbf{Z}\mathbf{Z}' + \sigma_e^2\mathbf{I}_n,$$

so

$$\begin{aligned}Ev_n &= E\left[(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'(\mathbf{Z}\mathbf{u} + \mathbf{e})\right] \\ &= \mathrm{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'(\sigma_u^2\mathbf{Z}\mathbf{Z}' + \sigma_e^2\mathbf{I}_n)\right] \\ &= \sigma_e^2\mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\right] + \sigma_u^2\mathrm{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\right] \\ &\leq (\sigma_e^2 + N_{\max}\sigma_u^2)\mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\right]. \qquad (2.10)\end{aligned}$$

By Theorem 2.1,

$$\mathbf{\Sigma}_{\mathbf{x},1}^{(n)} - m^{-\delta}\mathbf{I}_{p_n} \leq m^{-1}\mathbf{X'}_n\mathbf{X}_n \leq \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} + m^{-\delta}\mathbf{I}_{p_n}.$$

Moreover, the matrix $\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}$ satisfies

$$m^*\mathbf{I}_{p_n} \leq \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \leq N_{\max}M^*\mathbf{I}_{p_n},$$

therefore

$$(m^* - m^{-\delta})\mathbf{I}_{p_n} \leq m^{-1}\mathbf{X'}_n\mathbf{X}_n \leq (N_{\max}M^* + m^{-\delta})\mathbf{I}_{p_n}$$

and by Corollary 1.3, $m^{-1}\mathbf{X'}_n\mathbf{X}_n$ is positive definite when m is large enough. So the smallest eigenvalue of $m^{-1}\mathbf{X'}_n\mathbf{X}_n$ is bounded below:

$$\lambda_{\min}(m^{-1}\mathbf{X'}_n\mathbf{X}_n) \geq m^* - m^{-\delta},$$

and

$$\begin{aligned} \mathrm{tr}\left[(\mathbf{X'X})^{-1}\right] &= \frac{1}{m}\mathrm{tr}\left[\left(\frac{\mathbf{X'X}}{m}\right)^{-1}\right] \leq \frac{1}{m}p_n\lambda_{\min}^{-1}(m^{-1}\mathbf{X'}_n\mathbf{X}_n) \\ &\leq p_n/[m(m^* - a_m)] = O(p_n/n). \end{aligned} \tag{2.11}$$

In conclusion,

$$P\left[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \frac{Kp_n}{\sqrt{n}}\right] \leq O(p_n/n \cdot n/p_n^2) = O(p_n^{-1}) \to 0,$$

implying $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(p_n/\sqrt{n})$. □

Remark: Therefore, even though the model is misspecified, the LS estimator of the coefficients derived from the working model is still consistent. Theorem 2.2 not only proves that the LS estimator asymptotically converges to $\boldsymbol{\beta}_0$, it also gives the rate of consistency in probability. From the proof we can see that the result holds as long as $\theta < 1/2$. □

2.2.2 The Asymptotic Variance of $\hat{\boldsymbol{\beta}}_n$

It is obvious that under the specification (2.1) when both $\mathbf{u}$ and $\mathbf{e}$ are normal vectors conditionally given $\mathbf{X}$:

$$\mathbf{y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_0, \mathbf{V}),$$

where $\mathbf{V} = \sigma_u^2\mathbf{Z}\mathbf{Z}' + \sigma_e^2\mathbf{I}$, and

$$\hat{\boldsymbol{\beta}}_n|\mathbf{X} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n}),$$

where

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

In other words, under the true model, both $\mathbf{y}$ and $\hat{\boldsymbol{\beta}}_n$ are normal conditional on $\mathbf{X}$. Since

$$m^{-1}\mathbf{X}'\mathbf{V}\mathbf{X} = m^{-1}(\sigma_u^2\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X} + \sigma_e^2\mathbf{X}'\mathbf{X}) = m^{-1}\left(\sigma_u^2\sum_{i=1}^{m} n_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i + \sigma_e^2\sum_{i=1}^{m} n_i\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right),$$

by Theorem 2.1 we get

$$\left\|m^{-1}\mathbf{X}'\mathbf{V}\mathbf{X} - \left(\sigma_u^2\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)\right\| \xrightarrow{p} 0,$$

and

$$\left\|m\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n} - \left[\sigma_u^2\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)}\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} + \sigma_e^2\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right]\right\| \xrightarrow{p} 0. \tag{2.12}$$

Therefore, the asymptotic variance of $\hat{\boldsymbol{\beta}}_n$ is (m^{-1} multiplied by)

$$\boldsymbol{\Sigma}_A = \sigma_u^2\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)}\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} + \sigma_e^2\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}. \tag{2.13}$$

2.2.3 Asymptotic Normality of $\hat{\boldsymbol{\beta}}_n$

In this section we are going to derive the asymptotic normality of $\sqrt{m}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$.

Theorem 2.3 *Under Assumptions 1.2–1.4 and 2.1, if* $r > 5\theta/[(2(1-3\theta)]$, *then for any sequence of unit vectors* $\mathbf{v}_n \in \mathbf{R}^{p_n}$,

$$\sqrt{m}\sigma_{\mathbf{v}_n}^{-1}\mathbf{v}_n'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to \mathcal{N}(0,1) \tag{2.14}$$

where $\sigma_{\mathbf{v}_n}^2 = \mathbf{v}_n'\boldsymbol{\Sigma}_A\mathbf{v}_n$ *and* $\boldsymbol{\Sigma}_A$ *is defined in (2.13).*

Proof: Let the $p_n \times 1$ vectors $\boldsymbol{\zeta}_i$ be defined as $\boldsymbol{\zeta}_i \equiv \sum_{j=1}^{n_i}(y_{ij} - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\tilde{\mathbf{x}}_i'$ and note that

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\tilde{\mathbf{x}}_i' = \mathbf{X}_n'(\mathbf{Z}\mathbf{u} + \mathbf{e}).$$

Then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= m^{-1}(m^{-1}\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Z}\mathbf{u} + \mathbf{e}) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\tilde{\mathbf{x}}_i' \\
&\quad + \left[(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right]\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\tilde{\mathbf{x}}_i' \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\zeta}_i + \left[(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right]\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\zeta}_i.
\end{aligned}$$

To prove (2.14), since $\sigma_{\mathbf{v}_n}$ is uniformly bounded away from 0 and ∞, we only need to have:

- $m^{-1/2}\mathbf{v}_n'\left[(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right]\sum_{i=1}^{m}\boldsymbol{\zeta}_i = o_p(1)$.
- The sequence $\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\zeta}_i$ satisfy the Lyapunov condition for any sequence of unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$.

Actually, if $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}$ denotes the covariance-variance matrix of $\boldsymbol{\zeta}_i$, and since $(M^*)^{-1}\mathbf{I}_{p_n} \leq \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} \leq (m^*)^{-1}\mathbf{I}_{p_n}$, the vector

$$\mathbf{v}_n^* = \frac{\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n}{\left\|\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right\|}$$

is also a unit vector in $\mathbf{R}^{p_n}$. If we can prove that

$$(\mathbf{v}_n'\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\mathbf{v}_n)^{-1/2}\mathbf{v}_n'\frac{1}{\sqrt{m}}\sum_i \boldsymbol{\zeta}_i \to \mathcal{N}(0,1)$$

for any sequence of unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$, then

$$\begin{aligned}
& \left[\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right]^{-1/2}\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\right)^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i \\
= & \left[\left\|\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right\|\mathbf{v}_n^{*\prime}\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\mathbf{v}_n^*\left\|\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right\|\right]^{-1/2}\left\|\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right\|\mathbf{v}_n^{*\prime}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i \\
= & \left(\mathbf{v}_n^{*\prime}\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\mathbf{v}_n^*\right)^{-1/2}\mathbf{v}_n^{*\prime}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i \to \mathcal{N}(0,1).
\end{aligned}$$

This indicates that the sequence $\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\zeta}_i$ satisfies the Lyapunov condition (Definition 1.13 for any sequence of unit vectors $\mathbf{v}_n$ if the sequence $\mathbf{v}_n'\boldsymbol{\zeta}_i$ satisfies the Lyapunov condition for any sequence of unit vectors $\mathbf{v}_n$.

Hence, we are going to prove instead that

1. $m^{-1/2}\mathbf{v}_n'\left[(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right]\sum_{i=1}^{m}\boldsymbol{\zeta}_i = o_p(1)$.

2. For any sequence of unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$, the sequence $\mathbf{v}_n'\boldsymbol{\zeta}_i$ satisfies the Lyapunov condition for central limit theorem.

As stated in Shiryaev [26], the Lyapunov condition is a sufficient condition for Lyapunov central limit theorem.

Let $e_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$ be the average of e_{ij} in the i^{th} cluster. Then given $\tilde{\mathbf{x}}_i$ and n_i, the random variable $(u_i + e_{i\cdot})$ is normally distributed with mean 0 and variance $(\sigma_u^2 + \sigma_e^2/n_i)$, and therefore has finite $(4r)^{th}$ moment. Therefore $\boldsymbol{\zeta}_i$ are iid $p_n \times 1$ random vectors with mean zero and

$$\begin{aligned}
E\left|\boldsymbol{\zeta}_{ik}\right|^{4r} &= E\left|\sum_{j=1}^{n_i}(y_{ij} - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\tilde{\mathbf{x}}_{ik}\right|^{4r} \\
&\leq ME\left|n_i\tilde{\mathbf{x}}_{ik}\right|^{4r} \leq C < \infty \qquad (2.15)
\end{aligned}$$

uniformly over k. Since

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\zeta}_i\right\| = \frac{1}{m}\sqrt{\sum_{k=1}^{p_n}\left(\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right)^2} \le \sqrt{p_n}\max_{1\le k\le p_n}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right|,$$

for any $0 < \delta < \frac{r(1-2\theta)-2\theta}{2r}$, and any constant $K > 0$, by Chebyshev Inequality, Burkholder's Inequality and (2.15),

$$\begin{aligned}
& P\left[\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i\right\| > Kn^{\delta-\varepsilon}\right] \\
\le\ & P\left[\sqrt{p_n m}\max_{1\le k\le p_n}\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right| > Kn^{\delta-\varepsilon}\right] \\
\le\ & p_n P\left[\left|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right|^{4r} > \left(Kn^{\delta-\varepsilon}m^{-1/2}p_n^{-1/2}\right)^{4r}\right] \\
\le\ & M_r p_n m^{-2r}K^{-4r}m^{2r}n^{-4r\delta+4r\varepsilon}p_n^{2r} \\
=\ & O(p_n^{1+2r}n^{-4r\delta+4r\varepsilon}) \to 0, \qquad (2.16)
\end{aligned}$$

if $\varepsilon > 0$ is sufficiently small and $4r\delta > \theta(1+2r)$. Since $\delta < \frac{r(1-2\theta)-2\theta}{2r}$ is arbitrary, when

$$\frac{\theta(1+2r)}{4r} < \frac{r(1-2\theta)-2\theta}{2r},$$

or $r > 5\theta/[2(1-3\theta)]$, we have

$$\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i\right\| = O_p(n^{\delta-\varepsilon})$$

for ε sufficiently small and any δ such that

$$\frac{\theta(1+2r)}{4r} < \delta < \frac{r(1-2\theta)-2\theta}{2r}.$$

Note that

$$\frac{5\theta}{2(1-3\theta}) > \frac{2\theta}{1-2\theta}$$

for all $0 < \theta < 1/3$, so $r > 5\theta/[2(1-3\theta)]$ automatically means that $r > 2\theta/(1-2\theta)$ for $\theta < 1/4$.

This is to say that for any unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$, there exists $\delta > 0$ such that $\left\|(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right\| = O_p(n^{-\delta})$ and $\left\|m^{-1/2}\sum_{i=1}^m \boldsymbol{\zeta}_i\right\| = O_p(n^{\delta-\varepsilon})$.

Therefore

$$\begin{aligned}
& \sqrt{m}\mathbf{v}_n' \left[(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right] \frac{1}{m}\sum_{i=1}^m \boldsymbol{\zeta}_i \\
\leq & \ \|\mathbf{v}_n\| \left\|(m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right\| \cdot \left\|\frac{1}{\sqrt{m}}\sum_{i=1}^m \boldsymbol{\zeta}_i\right\| \\
\leq & \ O_p(n^{-\delta}n^{\delta-\varepsilon}) \\
= & \ o_p(1) \qquad (2.17)
\end{aligned}$$

The first condition is proved.

Now we need to check the Lyapunov condition in Definition 1.13 for each sequence $m^{-1/2}\mathbf{v}_n'\sum_{i=1}^m \boldsymbol{\zeta}_i$. In other words, if $\sigma_{n,\mathbf{v}_n}^2 = \sum_{i=1}^m \mathrm{var}[\mathbf{v}_n'\boldsymbol{\zeta}_i]$, then we need to prove that $\sigma_{n,\mathbf{v}_n}$ is bounded away from zero for any unit vector $\mathbf{v}_n$ and

$$\sum_{i=1}^m \frac{E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^3}{\sigma_{n,\mathbf{v}_n}^3} \to 0.$$

First of all, by Chebyshev inequality, triangular inequality and (2.15),

$$\begin{aligned}
E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^3 &= E\left|\sum_{k=1}^{p_n} \mathbf{v}_{nk}\boldsymbol{\zeta}_{ik}\right|^3 \\
&\leq E[\sum_{k=1}^{p_n} \mathbf{v}_{nk}^2 \sum_{k=1}^{p_n} \boldsymbol{\zeta}_{ik}^2]^{3/2} \\
&= \left\|\sum_{k=1}^{p_n} \boldsymbol{\zeta}_{ik}^2\right\|_{3/2}^{3/2} \\
&\leq \left(\sum_{k=1}^{p_n} \|\boldsymbol{\zeta}_{ik}^2\|_{3/2}\right)^{3/2} \\
&\leq p_n^{3/2} \max_{1\leq k\leq p_n} E|\boldsymbol{\zeta}_{ik}|^3
\end{aligned}$$

$$\leq \quad p_n^{3/2} C. \tag{2.18}$$

Moreover,

$$\begin{aligned} E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^2 &= E\left[E\left[\left(\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\tilde{\mathbf{x}}_i\mathbf{v}_n\right)^2 |u_i, n_i, \tilde{\mathbf{x}}_i\right]\right] \\ &\geq E\left[E\left[\left(\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0-u_i)\tilde{\mathbf{x}}_i\mathbf{v}_n\right)^2 |u_i, n_i, \tilde{\mathbf{x}}_i\right]\right] \\ &\geq \sigma_e^2 E(\tilde{\mathbf{x}}_i\mathbf{v}_n)^2 \\ &\geq \sigma_e^2 \mathbf{v}_n' E\left[\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right]\mathbf{v}_n \\ &\geq \sigma_e^2 m^* \end{aligned} \tag{2.19}$$

Therefore,

$$\sum_{i=1}^{m} \frac{E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^3}{\sigma_{n,\mathbf{v}_n}^3} \leq \frac{m p_n^{3/2} C}{m^{3/2}(\sigma_e^2 m^*)^{3/2}} \to 0,$$

and for any sequence of unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$,

$$\left[\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right]^{-1/2}\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i \to \mathcal{N}(0,1). \tag{2.20}$$

Finally,

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\zeta}} = E[\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i'] &= E\left[\mathbf{x}_i'\left(\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\right)^2\mathbf{x}_i\right] \\ &= E\left[\mathbf{x}_i' E\left[\left(\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\right)^2 |u_i, n_i, \tilde{\mathbf{x}}_i\right]\mathbf{x}_i\right] \\ &= E\left[\mathbf{x}_i'(n_i^2\sigma_u^2 + n_i\sigma_e^2)\mathbf{x}_i\right] \\ &= \sigma_u^2\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}, \end{aligned} \tag{2.21}$$

which implies for any sequence of unit vectors $\mathbf{v}_n$,

$$\left[\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\mathbf{v}_n\right]^{-1/2}\mathbf{v}_n'\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{\zeta}_i \to \mathcal{N}(0,1)$$

by Lyapunov CLT. By the definition of $\mathbf{\Sigma}_A$ in (2.13), we can see that

$$\left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} \mathbf{\Sigma}_{\boldsymbol{\zeta}} \left(\mathbf{\Sigma}_{\mathbf{x}}^{(n)}\right)^{-1} = \mathbf{\Sigma}_A$$

and therefore for any unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$,

$$\begin{aligned}
&\sqrt{m}(\mathbf{v}_n'\mathbf{\Sigma}_A\mathbf{v}_n)^{-1/2}\mathbf{v}_n'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&= \left[\mathbf{v}_n' \left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} \mathbf{\Sigma}_{\boldsymbol{\zeta}} \left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} \mathbf{v}_n\right]^{-1/2} \mathbf{v}_n' \left(\mathbf{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \boldsymbol{\zeta}_i + o_p(1) \\
&\overset{(2.20)}{\to} \mathcal{N}(0,1), \qquad (2.22)
\end{aligned}$$

which, by Definition 1.12, means that $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ is asymptotically normal. □

Remark: The asymptotic normality discussed in Theorem 2.3 is very strong, since it implies that each entry of $\hat{\boldsymbol{\beta}}_n$ is asymptotically normal. For this result to hold, it is clear that θ has to be less than 1/3. This means that the rate at which p_n must grow is smaller for asymptotic normality than for consistency of $\hat{\boldsymbol{\beta}}_n$. □

2.2.4 The Variance Estimators

The estimator of the variance σ_0^2 under the working model (2.3) is usually the mean Residual Sum of Squares (RSS):

$$s^2 = (n - p_n)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Let

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Then clearly

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H},$$

so the symmetric matrices $\mathbf{H}$ and $(\mathbf{I}-\mathbf{H})$ are idempotent. Therefore

$$s^2 = (n-p_n)^{-1}(\mathbf{y}-\mathbf{H}\mathbf{y})'(\mathbf{y}-\mathbf{H}\mathbf{y}) = (n-p_n)^{-1}(\mathbf{Z}\mathbf{u}+\mathbf{e})'(\mathbf{I}-\mathbf{H})(\mathbf{Z}\mathbf{u}+\mathbf{e}). \quad (2.23)$$

When the model is correctly specified, this estimator is consistent ($s^2 \xrightarrow{p} \sigma_0^2$). The following theorem establishes the limit of s^2 in probability when $n \to \infty$.

Theorem 2.4 *The sample variance s^2 under the working model converges to $(\sigma_u^2 + \sigma_e^2)$ in probability, where the probability is taken under the true model.*

Proof: To see the asymptotic limit of s^2 under the true model (2.1), note that for any random variable T, $P(|T-ET| > \epsilon) \le \mathrm{var}T/\epsilon^2$. The first step is to find the expectation and variance of s^2. First,

$$\begin{aligned} Es^2 &= (n-p_n)^{-1}E\left[(\mathbf{Z}\mathbf{u}+\mathbf{e})'(\mathbf{I}-\mathbf{H})(\mathbf{Z}\mathbf{u}+\mathbf{e})\right] \\ &= (n-p_n)^{-1}\mathrm{tr}((\mathbf{I}-\mathbf{H})\mathbf{V}) \\ &\overset{(2.5)}{=} \sigma_e^2 + \frac{n-\mathrm{tr}(\mathbf{H}\mathbf{Z}\mathbf{Z}')}{n-p_n}\sigma_u^2, \end{aligned} \quad (2.24)$$

where the expectation is taken conditional on $\mathbf{X}$ under the true model. For nonnegative definite matrix $\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'$ and diagonal matrix $\mathbf{Z}'\mathbf{Z}$, by Proposition A.1

$$\begin{aligned} \mathrm{tr}[\mathbf{H}\mathbf{Z}\mathbf{Z}'] &= \mathrm{tr}[\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{Z}'\mathbf{Z}\mathbf{Z}'\mathbf{Z}] \\ &\le N_{\max}\mathrm{tr}[\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{Z}'\mathbf{Z}] \\ &= N_{\max}\mathrm{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = N_{\max}p_n. \end{aligned} \quad (2.25)$$

So $\mathrm{tr}(\mathbf{H}\mathbf{Z}\mathbf{Z}') = o(n)$ since $N_{\max} < \infty$ and $p_n = o(n)$. Therefore, when $n \to \infty$, $Es^2 \to \sigma_u^2 + \sigma_e^2$.

Under normal-distribution assumptions on $\mathbf{u}$ and $\mathbf{e}$ in (2.1), the variance of s^2 is equal to

$$\begin{aligned}
\mathrm{var}((\mathbf{Zu}+\mathbf{e})'(\mathbf{I}-\mathbf{H})(\mathbf{Zu}+\mathbf{e})) &= 2\mathrm{tr}[((\mathbf{I}-\mathbf{H})\mathbf{V})^2] \\
&= 2\mathrm{tr}\left[((\mathbf{I}-\mathbf{H})(\sigma_{\mathbf{u}}^{\mathbf{2}}\mathbf{ZZ}'+\sigma_{\mathbf{e}}^{\mathbf{2}}\mathbf{I}))^2\right] \\
&= 2\sigma_u^4\mathrm{tr}\left[(\mathbf{I}-\mathbf{H})\mathbf{ZZ}'(\mathbf{I}-\mathbf{H})\mathbf{ZZ}'\right] \\
&+ 4\sigma_u^2\sigma_e^2\mathrm{tr}\left[(\mathbf{I}-\mathbf{H})\mathbf{ZZ}'\right]+2\sigma_e^4\mathrm{tr}(\mathbf{I}-\mathbf{H}) \\
&\overset{(2.6)}{=} 2\sigma_u^4\left[\mathrm{tr}(\mathbf{ZZ}'\mathbf{ZZ}')-2\mathrm{tr}(\mathbf{HZZ}'\mathbf{ZZ}')+\mathrm{tr}(\mathbf{HZZ}'\mathbf{HZZ}')\right] \\
&+ 2\sigma_e^4(n-p_n)+4\sigma_u^2\sigma_e^2(n-o(n)). \qquad (2.26)
\end{aligned}$$

Note that

$$\frac{\mathrm{tr}(\mathbf{ZZ}'\mathbf{ZZ}')}{n}=\frac{\sum_{i=1}^m n_i^2}{n}=\frac{\sum_i n_i^2}{m}\cdot\frac{m}{\sum_i n_i}\to\frac{N_2}{N_1}<\infty,$$

$$\mathrm{tr}[(\mathbf{HZZ}'\mathbf{ZZ}')=\mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathrm{diag}(n_1^3,n_2^3,\cdots,n_m^3)\tilde{\mathbf{X}}]\le N_{\max}^2 p_n=o(n),$$

and

$$\begin{aligned}
\mathrm{tr}[\mathbf{HZZ}'\mathbf{HZZ}'] &= \mathrm{tr}[\mathbf{Z}\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{Z}'\mathbf{ZZ}'\mathbf{Z}\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{Z}'\mathbf{ZZ}'] \\
&\le N_{\max}\mathrm{tr}[\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{Z}'\mathbf{ZZ}'\mathbf{Z}] \\
&\le N_{\max}^2\mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\mathbf{Z}'\mathbf{Z}\tilde{\mathbf{X}}'] \\
&\le N_{\max}^2 p_n=o(n). \qquad (2.27)
\end{aligned}$$

Thus the variance of s^2 satisfies

$$\mathrm{var}(s^2)=\frac{2\sigma_u^4(O(n)+o(n))+4\sigma_u^2\sigma_e^2(n-o(n))+2\sigma_e^4(n-p_n)}{(n-p_n)^2}=O(n^{-1}).$$

From the above calculation we see that$Es^2\to\sigma_u^2+\sigma_e^2$ and $\mathrm{var}(s^2)\to 0$ as $n\to\infty$, which implies $s^2\xrightarrow{p}\sigma_u^2+\sigma_e^2$. □

Remark: The sample variance s^2 derived from the working model is biased. This is due to the failure of the working model to account for the variability introduced by the random intercept at the cluster level. Theorem 2.4 also bounds the difference between s^2 and $(\sigma_u^2 + \sigma_e^2)$ in probability: $s^2 = \sigma_u^2 + \sigma_e^2 + O_p(n^{-1/2+\varepsilon})$ where ε is an arbitrarily small positive number. This follows easily from Chebyshev's Inequality and the fact that $\text{var}(s^2) = O(n^{-1})$. □

Under the working model,

$$\hat{\boldsymbol{\beta}}_n | \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}), \tag{2.28}$$

and Theorem 2.4 gives an estimator under the working model for the variance in (2.28).

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}_n) = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

There is also a robust choice of variance estimator, that is, an estimator valid under the true model. Let $l(\boldsymbol{\theta})$ be the log-likelihood of the working model, that is,

$$\begin{aligned} l(\boldsymbol{\theta}_n) &= \sum_i \sum_j l_{ij}(\boldsymbol{\theta}_n) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n)}{2\sigma_0^2} \\ &= -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{\sum_{i=1}^m \sum_{j=1}^{n_i}(y_{ij} - \mathbf{x}_i\boldsymbol{\beta}_n)^2}{2\sigma_0^2}. \end{aligned} \tag{2.29}$$

with $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_n, \sigma_0^2)'$. Recall that $\boldsymbol{\nabla}_t f$ denotes the gradient of a function f with respect to t and $(\boldsymbol{\nabla}_t^{\otimes 2} f)$ the Hessian of f with respect to t. We define

$$\mathbf{A}_n(\boldsymbol{\theta}_n) = -m^{-1}\boldsymbol{\nabla}_{\boldsymbol{\theta}_n}^{\otimes 2} l(\boldsymbol{\theta}_n) \tag{2.30}$$

and

$$\mathbf{B}_n(\boldsymbol{\theta}_n) = m^{-1}\sum_i \left(\boldsymbol{\nabla}_{\boldsymbol{\theta}_n} \sum_{j=1}^{n_i} l_{ij}(\boldsymbol{\theta}_n)\right)^{\otimes 2}$$

$$= \quad m^{-1} \sum_i \begin{pmatrix} \sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_{ij}\boldsymbol{\beta}_n)\mathbf{x}'_{ij}/\sigma_0^2 \\ -1/(2\sigma_0^2) + (y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}_n)^2/(2\sigma_0^4) \end{pmatrix}^{\otimes 2}$$

Then the robust variance estimator for $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_0^2)'$ is defined as

$$\hat{\text{var}}_R(\hat{\boldsymbol{\theta}}_n) = \mathbf{A}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)\mathbf{A}_n^{-1}(\hat{\boldsymbol{\theta}}_n), \tag{2.31}$$

where $\hat{\sigma}_0^2 = [(n - p_n)/n]s^2$ is the Maximum Likelihood Estimator (MLE) for σ_0^2 under the working model. This robust variance estimator, also called the "sandwich" variance estimator is a $(p_n + 1) \times (p_n + 1)$ matrix. This is Huber's [13] and White's [32] definition for the robust variance estimator. By the definition of $\hat{\boldsymbol{\beta}}_n$ and $\hat{\sigma}^2$, we get:

$$\mathbf{A}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{m} \begin{pmatrix} (\mathbf{X}'\mathbf{X})/\hat{\sigma}_0^2 & \mathbf{0} \\ \mathbf{0} & 2n/\hat{\sigma}_0^2 \end{pmatrix}.$$

Let $\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)_{11}$ be the $p_n \times p_n$ matrix at the upper-left block of $\mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)$, then the upper left block of $\hat{\text{var}}_R(\hat{\boldsymbol{\theta}}_n)$ gives an estimator for the variance of $\hat{\boldsymbol{\beta}}_n$, and it is equal to

$$\hat{\text{var}}_R(\hat{\boldsymbol{\beta}}_n) = \left(\frac{1}{m}\mathbf{X}'\mathbf{X}\right)^{-1} \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{x}}'_i\right)^{\otimes 2} \left(\frac{1}{m}\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{2.32}$$

We have

Theorem 2.5 *Under Assumptions 1.2-1.4 and 2.1, if further we have* $r > 1/[2(1 - 3\theta)]$, *then*

$$\left\| \hat{var}_R(\hat{\boldsymbol{\beta}}_n) - \boldsymbol{\Sigma}_A \right\| \xrightarrow{p} 0,$$

i.e., the "sandwich" variance estimator converges to the true variance $\boldsymbol{\Sigma}_A$ *in probability.*

Proof: Since

$$\left\| (m^{-1}\mathbf{X}'\mathbf{X})^{-1} - \left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1} \right\| \xrightarrow{p} 0$$

and

$$(M^* + a_n)^{-1} \leq \|(m^{-1}\mathbf{X}'\mathbf{X})^{-1}\| \leq (m^* + a_n)^{-1},$$

by the definition of $\mathbf{\Sigma}_A$ in (2.13), it suffices to show that

$$\left\| m^{-1} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_n) \tilde{\mathbf{x}}_i' \right)^{\otimes 2} - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right) \right\| \xrightarrow{p} 0. \tag{2.33}$$

To prove (2.33), note that

$$\begin{aligned} & \left\| m^{-1} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_n) \tilde{\mathbf{x}}_i' \right)^{\otimes 2} - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right) \right\| \\ \leq & \frac{1}{m} \left\| \sum_{i=1}^{m} \left[\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_n) \right]^2 \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i - \sum_{i=1}^{m} \left[\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_0) \right]^2 \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i \right\| \\ & + \left\| \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_0) \right]^2 \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right) \right\|. \end{aligned}$$

Let the $p_n \times p_n$ matrix

$$\begin{aligned} \mathbf{M} &\equiv \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{j=1}^{n_i} (y_{ij} - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_0) \right]^2 \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right) \\ &= \frac{1}{m} \sum_{i=1}^{m} n_i^2 (u_i + e_{i\cdot})^2 \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right) \end{aligned}$$

where $e_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$ is the average of e_{ij}'s in the i^{th} cluster. Let

$$\mathbf{M}_{kl} = \frac{1}{m} \sum_{i=1}^{m} n_i^2 (u_i + e_{i\cdot})^2 \tilde{\mathbf{x}}_{ik} \tilde{\mathbf{x}}_{il} - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right)_{kl}$$

be the elements of $\mathbf{M}$ for $1 \leq k, l \leq p_n$, and let

$$\eta_i^{(kl)} \equiv n_i^2 (u_i + e_{i\cdot})^2 \tilde{\mathbf{x}}_{ik} \tilde{\mathbf{x}}_{il} - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right)_{kl},$$

then for each k and l, $\mathbf{M}_{kl}$ is the average of m iid random variables with

$$E\eta_i^{(kl)} = E(n_i^2 E(u_i + e_{i\cdot})^2 \tilde{\mathbf{x}}_{ik} \tilde{\mathbf{x}}_{il}) - \left(\sigma_u^2 \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2 \mathbf{\Sigma}_{\mathbf{x},1}^{(n)} \right)_{kl}$$

$$
\begin{aligned}
&= E\left[n_i^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}E\left((u_i+e_{i\cdot})^2|n_i\right)\right] - \left(\sigma_u^2\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)_{kl} \\
&= \sigma_u^2 E[n_i^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}] + \sigma_e^2 E[n_i\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}] - \left(\sigma_u^2\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)} + \sigma_e^2\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)_{kl} \\
&= 0. \qquad (2.34)
\end{aligned}
$$

And since given n_i, $(u_i + e_{i\cdot}) \sim \mathcal{N}(0, \sigma_u^2 + \sigma_e^2/n_i)$ and has finite $(4r)^{th}$ moment,

$$
E|n_i^2(u_i+e_{i\cdot})^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}|^{2r} = E\left[|n_i^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}|^{2r}E((u_i+e_{i\cdot})^{4r}|n_i)\right] \leq CE|n_i^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}|^{2r} \leq C^*
$$

for $C^* < \infty$ uniformly in k and l by Assumption 1.3. Therefore the matrix $\mathbf{M}$ consists of a sum of zero-mean random variables with uniformly bounded $(2r)^{th}$ moments. Following the same arguments as in Theorem 1.1, there exists $\delta_M > 0$ such that $\|\mathbf{M}\| \leq n^{-\delta_M} \to 0$ with probability approaching 1.

On the other hand,

$$
\begin{aligned}
&\frac{1}{m}\left\|\sum_{i=1}^m\left[\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_n)\right]^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i - \sum_{i=1}^m\left[\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\right]^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| \\
= &\frac{1}{m}\left\|\sum_{i=1}^m\left[\left(\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_n)\right)^2 - \left(\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\right)^2\right]\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| \\
= &\frac{1}{m}\left\|\sum_{i=1}^m\left[\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_n+y_{ij}-\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\right]\left[\sum_{j=1}^{n_i}(y_{ij}-\tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_n-y_{ij}+\tilde{\mathbf{x}}_i\boldsymbol{\beta}_0)\right]\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| \\
= &\frac{1}{m}\left\|\sum_{i=1}^m\left[2n_i(u_i+e_{\cdot}) + n_i\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)\right]n_i(\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n))\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| \\
\leq &\frac{2}{m}\left\|\sum_{i=1}^m n_i^2(u_i+e_{i\cdot})\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| + \frac{1}{m}\left\|\sum_{i=1}^m n_i^2(\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n))^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| \\
\leq &\max_i|\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)|\left\|\frac{2}{m}\sum_{i=1}^m n_i^2(u_i+e_{i\cdot})\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\| + \max_i|\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)|^2\left\|\frac{1}{m}\sum_{i=1}^m n_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right\|.
\end{aligned}
$$

Since $\left\|m^{-1}\sum_{i=1}^m n_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i - \boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)}\right\| \xrightarrow{p} 0$ and $\left\|\boldsymbol{\Sigma}_{\mathbf{x},2}^{(n)}\right\|$ is bounded, $\|m^{-1}\sum_{i=1}^m n_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\| = O_p(1)$. Therefore, we need to prove that

1. $\max_i|\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)|\,\|m^{-1}\sum_{i=1}^m n_i^2(u_i+e_{i\cdot})\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\| \xrightarrow{p} 0$;

2. $\max_i |\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n|^2 \xrightarrow{p} 0$.

Let the $p_n \times p_n$ matrix $\mathbf{W} \equiv m^{-1}\sum_{i=1}^m n_i^2(u_i + e_{i\cdot})\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i$, then each element of $\mathbf{W}$ is the average of m iid random variables with $E[\mathbf{W}_{kl}] = E[n_i^2(u_i + e_{i\cdot})\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i] = 0$ and

$$E|\mathbf{W}_{kl}|^{2r} = E[|n_i^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}|^{2r}E(|u_i + e_{i\cdot}|^{2r}|n_i)] \leq CE|n_i^2\tilde{\mathbf{x}}_{ik}\tilde{\mathbf{x}}_{il}|^{2r} \leq C^*$$

uniformly in k and l for $C^* < \infty$. Again $\mathbf{W}$ is a matrix whose elements are average of iid zero-mean random variables with uniformly finite $(2r)^{th}$ moments. Therefore

$$\|\mathbf{W}\| = \|m^{-1}\sum_{i=1}^m n_i^2(u_i + e_{i\cdot})\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\| \xrightarrow{p} 0 \tag{2.35}$$

with probability approaching 1.

By Theorem 1.2 and Theorem 2.2 for $r > 1/(2(1-3\theta))$, let $\delta_4 = (1 - 3\theta - 1/(2r))/2 > 0$, then

$$\begin{aligned}
\max_i |\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_n)|^2 &\leq \max_i \|\tilde{\mathbf{x}}_i\|^2\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 \\
&= O_p(p_n n^{\frac{1}{2r}+\delta_4}p_n^2 n^{-1}) \\
&= O_p(n^{-\frac{1}{2}(1-3\theta-\frac{1}{2r})}) \to 0.
\end{aligned} \tag{2.36}$$

by (2.36) and (2.35), $\max_i |\tilde{\mathbf{x}}_i(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)|\|m^{-1}\sum_{i=1}^m n_i^2(u_i + e_{i\cdot})\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\| \xrightarrow{p} 0$. Finally,

$$\frac{1}{2(1-3\theta)} > \frac{2\theta}{1-2\theta}$$

whenever $\theta < 1/3$. The theorem is therefore proved. □

Remark: Theorem 2.5 tells us that when the model is misspecified, we can still get a consistent variance estimator for $\hat{\boldsymbol{\beta}}_n$ in the operator norm. Therefore, when $\theta < 1/3$ and r is large enough, the LS estimator derived from the working model is consistent, asymptotically normal, and has a consistent variance estimator.

□

2.3 Bonferroni-Adjusted Model Selection Procedure

Under the model misspecification indicated by (2.1) and (2.3), one quantity is of great interest to us: the expected number of extra regression coefficients to be tested significant in multiple hypothesis testing. The most commonly controlled quantity when testing multiple hypotheses is the experiment-wise error rate, which is the probability of yielding one or more false positives out of the p_n hypotheses tested:

$$P_{H_0}(\exists \text{ at least one false positive}) \leq \alpha. \tag{2.37}$$

Standard linear model theory and Analysis of Variance (ANOVA) treat the Multiple Comparison Procedure (MCP) with the statistics $\hat{\boldsymbol{\beta}}_n^{(k)}/\sqrt{\gamma_k s^2}$, where $\hat{\boldsymbol{\beta}}_n^{(k)}$ is the k^{th} entry of $\hat{\boldsymbol{\beta}}_n$, s^2 is the sample variance derived from the working model, and γ_k is the k^{th} diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. The coefficient $\boldsymbol{\beta}_k$ is said to be significant if and only if the absolute value of its standardized estimator is greater than some threshold decided by level α. Another Multiple Comparison Procedure often considered is that of Scheffé, which controls the error rate for any linear combination of the estimated coefficients to exceed a threshold. However, this procedure is not applicable for our model-selection procedures because we wish to select model terms one by one and not in linear combinations. Under either the working or the true model,

$$P_{H_0}(\exists \text{ at least one false positive}) = 1 - P_{H_0}(\frac{|\hat{\boldsymbol{\beta}}_n^{(k)}|}{\sqrt{\gamma_k s^2}} \leq t, \forall k = 1, \cdots, p_n). \tag{2.38}$$

By the Bonferroni Inequality,

$$P_{H_0}(\frac{[|\hat{\boldsymbol{\beta}}_n^{(k)}|}{\sqrt{\gamma_k s^2}} \leq t, \forall k = 1, \cdots, p_n) > 1 - \sum_{k=1}^{p_n} P_{H_0}(\frac{|\hat{\boldsymbol{\beta}}_n^{(k)}|}{\sqrt{\gamma_k s^2}} > t).$$

Under the null hypothesis $H_0 : \boldsymbol{\beta}_k = 0$, the standardized estimator of the coefficient is $t-$distributed, and (2.38) becomes

$$P_{H_0}(\exists \text{ at least one false positive}) \leq p_n P(|T| > t),$$

where T is $t-$distributed under the working model. Therefore, to have (2.37), it is sufficient to have $P(|T| > t) \leq \alpha/p_n$. Since under H_0 the standardized estimator converge to normal when n gets large, in large-sample asymptotics we usually use Z, a standard normal rather than T. Hence to control the experiment wise error rate at level α, we have $P(|Z| > t) = 2(1 - \Phi(t)) \leq \alpha$, where $\Phi(\cdot)$ is the cumulative density function of a standard normal variate. The threshold t could thus be determined under the working model:

$$t = \Phi^{-1}\left(1 - \frac{\alpha}{2p_n}\right). \tag{2.39}$$

This is derived from the Bonferroni Inequality, and is a rather stringent threshold to select significant variables, especially when $\hat{\boldsymbol{\beta}}_n^{(k)}$'s are correlated. A Bonferroni-Adjusted model selection has type I error controlled at level α. Our main interest lies in N_e, the expected number of extra variables that will be significant in a Bonferroni-Adjusted Model Selection due to the model misspecification. Ideally this should be controlled.

Theorem 2.6 *Suppose that the true model is specified as (2.1) and we are selecting a model according to the working model (2.3). Under Assumptions 1.2-1.4 and*

2.1, we use the statistic $\hat{\boldsymbol{\beta}}_n^{(k)}/\sqrt{\gamma_k s^2}$ *in the multiple hypothesis testing, then for* $r > 2\theta/(1-4\theta)$, *the Bonferroni-Adjusted Model Selection at level* α *with threshold* t *determined in equation (2.39) gives us a model with expected number of extra variables*

$$N_e \sim \sum_{k=p^*+1}^{p_n} r_k^{-1/2} \left(\frac{\alpha}{an^\theta}\right)^{r_k} (\pi\theta \log n)^{\frac{r_k-1}{2}}$$

where r_k *is the ratio of the* k^{th} *diagonal element of the two matrices* $(\sigma_u^2+\sigma_e^2)\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}$ *and* $\boldsymbol{\Sigma}_A$.

Proof: First of all, for any large positive number x (see proof in Appendix, Proposition B.2),

$$1-\Phi(x) = \frac{1}{\sqrt{2\pi}x} e^{-\frac{x^2}{2}}(1+O(x^{-2}))$$

Note that t is large because α is fixed but p_n is large. This means

$$\frac{\alpha}{2p_n} = \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}(1+O(t^{-2})). \tag{2.40}$$

Since $p_n = [an^\theta]$, taking logarithms on both sides of (2.40) yields

$$\log\alpha - \log(2a) - \theta\log n = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log t^2 - \frac{t^2}{2} + \log(1+O(t^{-2})).$$

Evidently the top order of t^2 is $\log n$, and t^{-2} is small, so that

$$\log(1+O(t^{-2})) = O(t^{-2}) = O((\log n)^{-1}).$$

Let $t^2 = 2\theta\log n + R_n$ where $R_n/\log n \to 0$; then (2.40) becomes

$$\log\left(\frac{\alpha\sqrt{2\pi}}{2a}\right) + \frac{1}{2}\log\left[(2\theta\log n)\left(1+\frac{R_n}{2\theta\log n}\right)\right] + \frac{R_n}{2} = O((\log n)^{-1}),$$

and

$$\begin{aligned}
&\log\left[(2\theta\log n)\left(1+\frac{R_n}{2\theta\log n}\right)\right] = \log(2\theta)+\log\log n+\log\left(1+\frac{R_n}{2\theta\log n}\right)\\
=\;& \log(2\theta)+\log\log n+\frac{R_n}{2\theta\log n}+O(R_n^2(\log n)^{-2})\\
=\;& \log(2\theta)+\log\log n+\frac{R_n}{2\theta\log n}+o(R_n/\log n) \qquad (2.41)
\end{aligned}$$

since we assume that $R_n/\log n \to 0$. Therefore (2.40) becomes

$$\begin{aligned}
R_n &= \left[\log\left(\frac{a^2}{\alpha^2\pi\theta}\right)-\log\log n+o(R_n/\log n\right]\left(1+\frac{1+o(1)}{2\theta\log n}\right)^{-1}\\
&= \left[\log\left(\frac{a^2}{\alpha^2\pi\theta}\right)-\log\log n+O((\log n)^{-1})\right]\left[1+O((\log n)^{-1})\right]\\
&= -\log\log n+\log\left(\frac{a^2}{\alpha^2\pi\theta}\right)+O\left(\frac{\log\log n}{\log n}\right). \qquad (2.42)
\end{aligned}$$

Therefore

$$t^2 = 2\theta\log n-\log\log n+\log\left(\frac{a^2}{\alpha^2\pi\theta}\right)+O\left(\frac{\log\log n}{\log n}\right). \qquad (2.43)$$

From (2.12) the true conditional variance given $\mathbf{X}$ for $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n}$. Therefore the true variance for the k^{th} entry of $\hat{\boldsymbol{\beta}}_n$, $\mathrm{var}(\hat{\boldsymbol{\beta}}_n^{(k)})$, is the k^{th} diagonal element of the matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n}$, $\left(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n}\right)_{kk}$. However, the working model uses $s^2(\mathbf{X}'\mathbf{X})^{-1}$ as the variance estimator of $\hat{\boldsymbol{\beta}}_n$, so that the k^{th} entry has an estimated variance of $\hat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_n^{(k)}) = s^2\left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}$. From Theorem 2.1, there exists $\delta_1 > 0$ such that

$$\left\|m\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n}-\boldsymbol{\Sigma}_A\right\| = O_p(n^{-\delta_1}), \quad \left\|m(\mathbf{X}'\mathbf{X})^{-1}-\left(\boldsymbol{\Sigma}_{\mathbf{x},1}^{(n)}\right)^{-1}\right\| = O_p(n^{-\delta_1}),$$

and from Theorem 2.4, $s^2 = (\sigma_u^2+\sigma_e^2)+O_p(n^{-1/2+\varepsilon_1})$ for a sufficiently small number $\varepsilon_1 > 0$. By the definition of operator norm, it is easy to show that the diagonal elements of a positive definite matrix are uniformly bounded by its operator norm,

therefore

$$m\left(\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}_n}\right)_{kk} = (\mathbf{\Sigma}_A) + O_p(n^{-\delta_1}), \text{ and } m\,(\mathbf{X}'\mathbf{X})^{-1}_{kk} = \left(\mathbf{\Sigma}^{(n)}_{\mathbf{x},1}\right)^{-1}_{kk} + O_p(n^{-\delta_1})$$

uniformly in k. The ratio of the true standard deviation to the estimated standard deviation is therefore

$$\sqrt{\frac{s^2\left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}{(\Sigma_{\hat{\boldsymbol{\beta}}_n})_{kk}}} = \sqrt{r_k}(1 + O_p(n^{-\delta_1})) \tag{2.44}$$

uniformly in k, and the constants in $O_p(n^{-\delta_1})$ do not depend on k.

The first p^* coefficients $\boldsymbol{\beta}^*_k$ are all nonzero under Assumption 1.1 and since $\hat{\boldsymbol{\beta}}_n$ is consistent, we expect that

$$P\left[\frac{\hat{\boldsymbol{\beta}}^{(k)}_n}{\sqrt{\gamma_k s^2}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2p_n}\right)\right] \tag{2.45}$$

is small enough to be ignored for our interest. Certain probabilities such as

$$P[(m^{-1}\mathbf{X}'\mathbf{X})^{-1} \geq C_* m^{1/3}], \quad P[s^2 \geq C_* m^{1/3}]$$

and $P[m\left(\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}_n}\right)_{kk} \leq C_* m^{-1/3}]$ need to be estimated for constant C_* to check that (2.45) is ignorable; but for our purposes we only need (2.45) to be of the order n^{-1} and the three probabilities mentioned can be checked with a little care and further calculations. We will limit our attention only to the later entries of $\hat{\boldsymbol{\beta}}_n$. For $k \geq p^* + 1$

$$\begin{aligned} & P\left[\frac{|\hat{\boldsymbol{\beta}}^{(k)}_n|}{\sqrt{\gamma_k s^2}} \geq \Phi^{-1}(1 - \frac{\alpha}{2p_n})\right] \\ = \; & P\left[\frac{|\hat{\boldsymbol{\beta}}_k|}{\sqrt{(\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}_n})_{kk}}} \geq \sqrt{\frac{s^2\gamma_k}{(\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}_n})_{kk}}}\Phi^{-1}(1 - \frac{\alpha}{2p_n})\right] \end{aligned}$$

$$
\begin{aligned}
&= P\left[\frac{|\hat{\boldsymbol{\beta}}_k|}{\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_n})_{kk}}} \geq \sqrt{r_k}(1+O(n^{-\delta_1}))\Phi^{-1}(1-\frac{\alpha}{2p_n})\right] \\
&= 2\left[1-\Phi(\sqrt{r_k}(1+O_p(n^{-\delta_1}))t)\right]. \qquad (2.46)
\end{aligned}
$$

Since $t = O(\sqrt{\log n})$, and the pdf of standard normal is bounded by $1/\sqrt{2\pi}$,

$$
\begin{aligned}
&\left|\Phi(\sqrt{r_k}(1+O_p(n^{-\delta_1}))t) - \Phi(\sqrt{r_k}t)\right| \\
\leq\ & 1/\sqrt{2\pi}\sqrt{r_k}tO_p(n^{-\delta_1}) \\
=\ & O_p(n^{-\delta_1}\sqrt{\log n})
\end{aligned}
$$

where the constants in $O_p(n^{-\delta_1}\sqrt{\log n})$ are uniform in k since all r_k are bounded by 1. It follows that

$$
\begin{aligned}
N_e &= \sum_{k=p^*+1}^{p_n} P\left[\frac{|\hat{\boldsymbol{\beta}}_k|}{\sqrt{\gamma_k s^2}} \geq \Phi^{-1}(1-\frac{\alpha}{2p_n})\right] \\
&= 2\sum_{k=p^*+1}^{p_n}\left[1-\Phi(\sqrt{r_k}(1+O_p(n^{-\delta_1}))\Phi^{-1}(1-\frac{\alpha}{2p_n}))\right] \\
&= 2\sum_{k=p^*+1}^{p_n}\left[1-\Phi(\sqrt{r_k}t)+O_p(n^{-\delta_1}\sqrt{\log n})\right]
\end{aligned}
$$

Using Proposition B.2 one more time, we get

$$
\begin{aligned}
1-\Phi(\sqrt{r_k}t) &= \frac{1}{\sqrt{2\pi r_k}t}e^{-r_k t^2/2}(1+O(t^{-2})) \\
&= \left[\frac{n^{-\theta r_k}(\sqrt{\log n})^{r_k}\left(\frac{a}{\alpha\sqrt{\pi\theta}}\right)^{-r_k}}{\sqrt{2\pi r_k}\sqrt{2\theta\log n-\log\log n}}(1+O(t^{-2}))\right] \\
&= O(n^{-\theta r_k}(\sqrt{\log n})^{r_k-1})(1+O(t^{-2})). \qquad (2.47)
\end{aligned}
$$

where the constants in $O(t^{-2})$ are uniformly bounded by 1 and therefore do not depend on k. Comparing $O_p(n^{-\delta_1}\sqrt{\log n})$ with $O(n^{-\theta r_k}(\sqrt{\log n})^{r_k-1})$, it is obvious that the O_p term is negligible if $\delta_1 > \theta$, since all r_k's are bounded by 1. Here $0 < \delta_1 <$

$(r(1-2\theta)-2\theta)/(2r)$ is an arbitrary number, so we only need $\theta < (r(1-2\theta)-2\theta)/(2r)$, or $r > 2\theta/(1-4\theta)$ for such a δ_1 to exist. Therefore,

$$\begin{aligned} N_e &= 2\sum_{k=p^*+1}^{p_n}[1-\Phi(\sqrt{r_k}t)] + O_p(n^{\theta-\delta_1}\sqrt{\log n}) \\ &= 2\sum_{k=p^*+1}^{p_n}\left[\frac{n^{-\theta r_k}(\sqrt{\log n})^{r_k}\left(\frac{a}{\alpha\sqrt{\pi\theta}}\right)^{-r_k}}{\sqrt{2\pi r_k}\sqrt{2\theta\log n-\log\log n}}(1+O(t^{-2}))\right] + O(n^{\theta-\delta_1}\sqrt{\log n}) \\ &\sim \sum_{k=p^*+1}^{p_n} r_k^{-1/2}\left(\frac{\alpha}{an^\theta}\right)^{r_k}(\pi\theta\log n)^{\frac{r_k-1}{2}} \end{aligned} \tag{2.48}$$

□

Remark: It is worth mentioning that the approximation in Theorem 2.6 requires that $\theta < 1/4$. Even though each O_p terms in (2.44) is small, when adding q_n of them, this could be non-negligible compared to the main term $(1-\Phi(\sqrt{r_k}t))$. Nevertheless we have established the asymptotic equality (2.48). □

Corollary 2.1 *If instead of $s^2(\mathbf{X}'\mathbf{X})^{-1}$ we use in hypothesis testing the sandwich variance estimator (2.32) for $\hat{\boldsymbol{\beta}}_n$, then under Assumptions 1.2-1.6 and 2.1, for*

$$r > \max\left(\frac{1}{2(1-3\theta)}, \frac{2\theta}{1-4\theta}\right),$$

the experiment-wise error rate is controlled at level α.

Proof: As proved in Theorem 2.5, the sandwich variance estimator converges to the true variance of $\hat{\boldsymbol{\beta}}_n$ in matrix norm. Therefore the ratio in (2.44) is equal to $(1+O_p(n^{-\delta_1}))$ and

$$\begin{aligned} N_e &= 2\sum_{k=p^*+1}^{p_n}\left[1-\Phi(t)+O_p(n^{-\delta_1}\sqrt{\log n})\right] \\ &= \sum_{k=p^*+1}^{p_n}\left[\frac{\alpha}{2p_n}+O_p(n^{-\delta_1}\sqrt{\log n})\right] \\ &= \alpha + O_p(n^{\theta-\delta_1}\sqrt{\log n}). \end{aligned} \tag{2.49}$$

Here $0 < \delta_1 < (r(1-2\theta)-2\theta)/(2r)$ is an arbitrary number. As long as δ_1 is also greater than θ, we the O_p term in (2.49) will dissappear and the experiment-wise error rate will be controlled at the right level α. For such a δ_1 to exist, we only need $\theta < (r(1-2\theta)-2\theta)/(2r)$, or $r > 2\theta/(1-4\theta) > 2\theta/(1-4\theta)$. □

Remark: Corollary 2.1 encourages the use of the sandwich variance estimator when we do multiple hypothesis testing. Although the sample variance s^2 estimates consistently the true variance of y_{ij}, under the working model, it is still problematic to base our statistical inference on the variance structure that the working model assumes. This can be corrected by a using robust variance estimator, the "sandwich" variance estimator defined in (2.32), which estimates consistently the variance structure of the estimated coefficients and thus can lead to much more accurate conclusions. □

The formula in Theorem 2.6 can be reduced if we make the following additional assumption:

Assumption 2.2 *The cluster sizes n_i are independent of $\mathbf{x}_i$.*

Then,

Corollary 2.2 *If in addition to the assumptions in Theorem 2.6 we further impose Assumption 2.2, then*

$$N_e \sim A\left(\frac{n^\theta}{\sqrt{\log n}}\right)^{1-\varrho},$$

where

$$\varrho = \frac{\sigma_u^2 + \sigma_e^2}{\sigma_e^2 + \frac{N_2}{N_1}\sigma_u^2}$$

and

$$A = \frac{\alpha^{\varrho}}{\sqrt{\varrho}} \left(\frac{a}{\sqrt{\pi\theta}} \right)^{1-\varrho}.$$

Proof: Under Assumption 2.2,

$$\mathbf{\Sigma}_{\mathbf{x},1}^{(n)} = N_1 \mathbf{\Sigma}_{\mathbf{x}}^{(n)}, \qquad \mathbf{\Sigma}_{\mathbf{x},2}^{(n)} = N_2 \mathbf{\Sigma}_{\mathbf{x}}^{(n)}.$$

So the ratio r_k is equal to

$$\varrho = \frac{\sigma_u^2 + \sigma_e^2}{\sigma_e^2 + \frac{N_2}{N_1}\sigma_u^2}$$

for all $k = p^* + 1, \cdots, p_n$. Therefore

$$N_e = 2 \sum_{k=p^*+1}^{p_n} \left[1 - \Phi(\sqrt{\varrho}(1 + O_p(n^{-\delta_1}))t) \right] = 2q_n \left[1 - \Phi(\sqrt{\varrho}(1 + O_p(n^{-\delta_1}))t) \right]$$

where the O_p terms are uniform in k. Using the approximation (2.43) and then Proposition B.2 again, we get:

$$\begin{aligned} N_e &= 2q_n \left[1 - \Phi(\sqrt{\varrho}t) + O_p(n^{-\delta_1}\sqrt{\log n}) \right] \\ &= an^{\theta}\varrho^{-1/2} \left(\frac{\alpha}{an^{\theta}} \right)^{\varrho} (\pi\varrho \log n)^{\frac{\varrho-1}{2}} + O_p(n^{\theta-\delta_1}\sqrt{\log n}) \\ &\sim A \left(\frac{n^{\theta}}{\sqrt{\log n}} \right)^{1-\varrho}. \end{aligned}$$

□

Remark: Theorem 2.6 assumes that $r > 2\theta/(1-4\theta)$, and this is not necessary for Corollary 2.2. The O_p term N_e is negligible as long as $\theta(1-\varrho) > \theta - \delta_1$. So δ_1 has to be greater than $\theta\varrho$, which means we need $r > 2\theta/(1-2\theta-2\varrho\theta)$. □

As we mentioned, Bonferroni Multiple Comparison Procedure is a rather stringent procedure, especially when the coefficient estimates are not independent. When the standard deviations of $\hat{\boldsymbol{\beta}}_n$ are correctly estimated, the expected number of extra

variables is controlled at a fixed small number α. But even this stringent procedure can not control the experiment-wise error rate at the expected level because of the model misspecification. The number ϱ defined in Corollary 2.2 is always strictly less than 1, since at least one of the clusters should have more than one observation. Therefore N_e goes to infinity as the sample size grows, and the rate is determined by ϱ and θ. The smaller ϱ is, meaning that $N_2/N_1\sigma_u^2 >> \sigma_e^2$, the faster N_e is going to infinity. On the other hand, the faster we allow p_n to grow with n, the faster N_e will grow with n as well.

2.4 Shao's GIC

Automatic model selection is a class of procedures to choose the optimal model by a certain criterion. There are many different selection criteria proposed in the literature. See Rao and Wu [21] for a detailed discussion. Asymptotic properties of these selection methods are also discussed by Shao [24]. The desirable asymptotic properties, according to Shao, are consistency and loss efficiency, where the final model is chosen to minimize or almost to minimize the criterion with high probability. Shao proposed a criterion GIC_{λ_n} that can specialize to several well-known model selection criteria, including AIC, BIC, Cross Validation, Mallows' C_p, etc. He also summarized in his paper the asymptotic behavior of various model selection procedures in different situations. Since Shao's GIC unifies a class of model selection methods, we are interested in studying the expected number of extra variables in the model selected by Shao's GIC, and in a way summarize what to expect for

various model selection methods.

2.4.1 Notations and definitions

We will adopt Shao's notations and definitions in this part of the chapter.

Let $\mathcal{A}_n$ be a class of subsets $\alpha \subset \{1, \cdots, p_n\}$ each of which represents the column-indices from $\mathbf{X}_n$ for a proposed model. The number of models in $\mathcal{A}_n$ is finite, but may depend on n. For each $\alpha \in \mathcal{A}_n$, let $p_n(\alpha)$ be the size of α (the number of parameters in model α), and $\mathbf{I}_n(\alpha)$ be the $p_n \times p_n(\alpha)$ matrix of zeros and ones such that $(\mathbf{I}_n(\alpha))_{ik} = I[i \in \alpha$, and i is the k^{th} element of $\alpha]$. Then $\mathbf{X}(\alpha) = \mathbf{X}\mathbf{I}_n(\alpha)$ is the design matrix for the model containing precisely the predictors with indices in α. The Least Square Estimator of $\mathbf{y}$ under model α is denoted by $\hat{\mathbf{y}}(\alpha) = \mathbf{H}_n(\alpha)\mathbf{y}$, where $\mathbf{H}_n(\alpha) = \mathbf{X}(\alpha)[\mathbf{X}'(\alpha)\mathbf{X}(\alpha)]^{-1}\mathbf{X}'(\alpha)$. Since we are more interested in the extra number of variables chosen by model selection under misspecification (2.1) and (2.3) than the effect of omitting one important covariate, we further assume for simplicity that $\forall\alpha, \quad \{1, 2, \cdots, p^*\} \subseteq \alpha \in \mathcal{A}_n$, i.e. the true fixed effects are always included in all models considered. The loss function is defined as follows:

$$L_n(\alpha) = \frac{\|\mu_n - \hat{\mu}_n(\alpha)\|^2}{n},$$

where μ_n is the expectation of $\mathbf{y}$ conditional on the covariates in the true model, $\hat{\mu}_n(\alpha) = \mathbf{H}_n(\alpha)\mathbf{y}$ is the LSE of $\mathbf{y}$ under model α, and $\|\cdot\|$ is the Euclidean norm. The goal is to minimize $L_n(\alpha)$ among all the models in $\mathcal{A}_n$, but since $L_n(\alpha)$ is not observable, we instead select the model that minimizes the GIC_{λ_n} criterion:

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\alpha)\|^2}{n} + \frac{\lambda_n p_n(\alpha)\hat{\sigma}^2(\alpha)}{n} \tag{2.50}$$

over $\alpha \in \mathcal{A}_n$ where $p_n(\alpha)$ denotes the number of variables in model α, $\hat{\mathbf{y}}(\alpha)$ is the LSE of $\mathbf{y}$ under model α, $\hat{\sigma}_n^2(\alpha)$ is an estimator of σ^2 , and $\{\lambda_n\}$ is a sequence of non-random numbers no less than 2 such that $\lambda_n/n \to 0$. Shao did not impose any restriction on the variance estimators $\hat{\sigma}^2(\alpha)$ in his definition of (2.50), but for our purposes we use the sample variance under model α, discussed in Section 2.4.3.

2.4.2 The Loss Function

The loss function $L_n(\alpha)$ is a criterion most model selection methods try to minimize, but whether choosing according to the loss function gives us the same model as choosing according to a specific model selection method needs to be checked carefully. The minimizer of the loss function can be obtained analytically only in special cases. We are now going to look at these cases and discuss the model that minimizes the loss function. The loss function is a measurement of the discrepancy between the estimated mean and the conditional mean of $\mathbf{y}$. With a cluster-structure for the data, we want to take into account the difference among clusters in the mean of $\mathbf{y}$. This means that aside from the fixed effects, we also should condition on the random effect $\mathbf{u}$. The loss function for model α is

$$\begin{aligned} L_n(\alpha) &= n^{-1}\|\mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{u} - \mathbf{H}_n(\alpha)\mathbf{y}\|^2 \\ &= n^{-1}\|(\mathbf{I}_n - \mathbf{H}_n(\alpha))\mathbf{Z}\mathbf{u} - \mathbf{H}_n(\alpha)\mathbf{e}\|^2 \\ &= n^{-1}\left(\mathbf{u}'\mathbf{Z}'(\mathbf{I}_n - \mathbf{H}_n(\alpha))\mathbf{Z}\mathbf{u} + \mathbf{e}'\mathbf{H}_n(\alpha)\mathbf{e}\right). \end{aligned} \tag{2.51}$$

In Section 2.3, we call a variable "extra" when it is not one of the first p^* fixed effects in the true model (2.1) but is tested significant in a hypothesis testing; we

use the same name for a variable that is not one of the true fixed effects in (2.1) but is chosen in the final model of a selection procedure. Let α_L denote the model that minimizes the loss function, and $p_n(\alpha_L)$ denote the total number of variables in α_L.

Balanced Data, Orthogonal Design

The first special case we will discuss assumes that

Assumption 2.3 $\forall 1 \leq i \leq m$, $n_i = b$ *and*

$$\mathbf{X}'\mathbf{X} = diag(\gamma_1^{-1}, \gamma_2^{-1}, \cdots, \gamma_{p_n}^{-1}).$$

This is the simplest case of all, a balanced-data, orthogonal design. For this case we have

Theorem 2.7 *Under the basic assumptions and Assumption 2.3,* $p_n(\alpha_L) - p^* \sim Binomial(q_n, P_a)$, *where*

$$P_a = \frac{1}{2} + \frac{1}{\pi} \arcsin\left(\frac{b\sigma_u^2 - \sigma_e^2}{b\sigma_u^2 + \sigma_e^2}\right).$$

Since we assume that $\forall \alpha \in \mathcal{A}_n, \{1, 2, \cdots, p^*\} \subseteq \alpha$, we only choose from the models that are "larger" than the minimal model, the model with only the p^* fixed effects. Suppose that the model α_s contains p_s variables, among which are the p^* fixed effects, and $p_s - p^*$ extra variables. The $n \times p_s$ design matrix of the model α_s is $\mathbf{X}(\alpha_s)$. Let $\mathbf{x}^{(k)}, \quad k = 1, 2, \cdots, p_n$ be the columns of $\mathbf{X}$. Suppose that the models α_{s_1} and α_{s_2} are those with design matrices $\mathbf{X}^{(k)} = (\mathbf{X}_s | \mathbf{x}^{(s_k)})$, for $k = 1, 2$. Evidently, α_{s_1} and α_{s_2} are obtained by adding the corresponding variables $\mathbf{x}^{(s_k)}$ into α_s. The following lemma represents the difference of α_{s_k} and α_s in the loss function.

Lemma 2.1

$$nL_n(\alpha_{s_k}) - nL_n(\alpha_s) = -b\sigma_u^2\eta_{s_k} + \sigma_e^2\zeta_{s_k},$$

where

$$\zeta_{s_k} \equiv \sigma_e^{-2}\gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{e} \sim \chi_1^2,$$

and

$$\eta_{s_k} \equiv b^{-1}\sigma_u^{-2}\gamma_{s_k}\mathbf{u}'\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\mathbf{u} \sim \chi_1^2$$

are independent random variables both following a χ_1^2 distribution.

Proof: Since $n_i = b, \quad \forall i = 1, 2, \cdots, m,$

$$\mathbf{Z}'\mathbf{Z} = b\mathbf{I}_n. \tag{2.52}$$

Let $\tilde{\mathbf{x}}^{(k)}$ be the k^{th} column of $\tilde{\mathbf{X}}$, then

$$\mathbf{x}^{(k)} = \mathbf{Z}\tilde{\mathbf{x}}^{(k)} \tag{2.53}$$

$$\mathbf{x}^{(k)\prime}\mathbf{x}^{(j)} = \gamma_k^{-1}\delta_{k,j}; \tag{2.54}$$

and

$$\tilde{\mathbf{x}}^{(k)\prime}\tilde{\mathbf{x}}^{(j)} = b^{-1}\mathbf{x}^{(k)\prime}\mathbf{x}^{(j)} = b^{-1}\gamma_k^{-1}\delta_{k,j}. \tag{2.55}$$

Note that

$$\begin{aligned}(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)})^{-1} &= \begin{pmatrix} \mathbf{X}'(\alpha_s)\mathbf{X}(\alpha_s) & \mathbf{X}'(\alpha_s)\mathbf{x}^{(s_k)} \\ \mathbf{x}^{(s_k)\prime}\mathbf{X}(\alpha_s) & \mathbf{x}^{(s_k)\prime}\mathbf{x}^{(s_k)} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{X}'(\alpha_s)\mathbf{X}(\alpha_s) & \mathbf{0} \\ \mathbf{0}' & \gamma_{s_k}^{-1} \end{pmatrix}^{-1}.\end{aligned}$$

Now, for $k = 1, 2$

$$\begin{aligned}
\mathbf{H}_n(\alpha_{s_k}) &= \mathbf{X}^{(k)}(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)})^{-1}\mathbf{X}^{(k)\prime} \\
&= (\mathbf{X}(\alpha_s)|\mathbf{x}^{(s_k)}) \begin{pmatrix} \mathbf{X}'(\alpha_s)\mathbf{X}(\alpha_s) & \mathbf{0} \\ \mathbf{0}' & \mathbf{x}^{(s_k)\prime}\mathbf{x}^{(s_k)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}'(\alpha_s) \\ \mathbf{x}^{(s_k)\prime} \end{pmatrix} \\
&= \mathbf{H}_n(\alpha_s) + \gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}. \qquad (2.56)
\end{aligned}$$

Therefore by (2.51)

$$\begin{aligned}
& nL_n(\alpha_{s_k}) - nL_n(\alpha_s) \\
= \;& \mathbf{u}'\mathbf{Z}'(\mathbf{H}_n(\alpha_s) - \mathbf{H}_n(\alpha_{s_k}))\mathbf{Z}\mathbf{u} + \mathbf{e}'(\mathbf{H}_n(\alpha_{s_k}) - \mathbf{H}_n(\alpha_s))\mathbf{e} \\
= \;& -\gamma_{s_k}\mathbf{u}'\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\mathbf{u} + \gamma_{s_k}\mathbf{e}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{e} \qquad (2.57)
\end{aligned}$$

Both $(\gamma_{s_k}\mathbf{u}'\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\mathbf{u})$ and $(\gamma_{s_k}\mathbf{e}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{e})$ are rank-1 quadratic forms in normal variables of $\sqrt{\gamma_{s_k}}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\mathbf{u}$ and $\sqrt{\gamma_{s_k}}\mathbf{x}^{(s_k)\prime}\mathbf{e}$. For $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_v)$, the quadratic form $\mathbf{v}'\mathbf{M}\mathbf{v}$ follows a central $\chi^2_{r_v}$ distribution if and only if the matrix $\mathbf{M}\mathbf{\Sigma}_v$ is idempotent, where the degrees of freedom $r_v = \text{rank}(\mathbf{M}\mathbf{\Sigma}_v)$. Details and proofs of these widely-known results can be found in Searle [23], Chap.2. Note that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_n)$, and

$$\gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime} \overset{(2.54)}{=} \gamma_{s_k}^2\mathbf{x}^{(s_k)}\gamma_{s_k}^{-1}\mathbf{x}^{(s_k)\prime} = \gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}.$$

Therefore the rank-1 matrix

$$\sigma_e^{-2}\gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\sigma_e^2\mathbf{I}_n = \gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}$$

is idempotent and

$$\zeta_{s_k} \equiv \sigma_e^{-2}\gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{e} \sim \chi_1^2,$$

or

$$\gamma_{s_k}\mathbf{e}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{e} = \sigma_e^2\zeta_{s_k},$$

where $\zeta_{s_k} \sim \chi_1^2$.

On the other hand, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2\mathbf{I}_m)$, and since

$$\gamma_{s_k}\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z} = b^2\gamma_{s_k}\tilde{\mathbf{x}}^{(s_k)}\tilde{\mathbf{x}}^{(s_k)\prime} = b\gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime},$$

so

$$\left(b^{-1}\gamma_{s_k}\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\right)^2 = \gamma_{s_k}\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime} = b^{-1}\gamma_{s_k}\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}.$$

Therefore the rank-1 matrix

$$b^{-1}\sigma_u^{-2}\gamma_{s_k}\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\sigma_u^2\mathbf{I}_m = b^{-1}\gamma_{s_k}\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}$$

is idempotent, which makes

$$\eta_{s_k} \equiv b^{-1}\sigma_u^{-2}\gamma_{s_k}\mathbf{u}'\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\mathbf{u} \sim \chi_1^2,$$

or

$$\gamma_{s_k}\mathbf{u}'\mathbf{Z}'\mathbf{x}^{(s_k)}\mathbf{x}^{(s_k)\prime}\mathbf{Z}\mathbf{u} = b\sigma_u^2\eta_{s_k},$$

where $\eta_{s_k} \sim \chi_1^2$. Finally, η_{s_k} and ζ_{s_k} are independent because $\mathbf{u}$ and $\mathbf{e}$ are. □

Lemma 2.2 *For two independent standard normal variables Z_1, Z_2, and two positive real numbers C_1, C_2,*

$$P(C_1|Z_1| > C_2|Z_2|) = \frac{1}{2} + \frac{1}{\pi}\arcsin\left(\frac{C_1^2 - C_2^2}{C_1^2 + C_2^2}\right).$$

Proof: First of all, let

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

the probability that W_1 and W_2 have the same sign is

$$P(W_1W_2 > 0) = P(W_1 < 0, W_2 < 0) + P(W_1 > 0, W_2 > 0) = 2P(W_1 > 0, W_2 > 0).$$

Note that the joint density of W_1 and W_2 is

$$f(w_1, w_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{w_1^2 - 2\rho w_1 w_2 + w_2^2}{2(1-\rho^2)}\},$$

so

$$\begin{aligned}
P(W_1W_2 > 0) &= 2\int_0^\infty \int_0^\infty f(w_1, w_2) dw_1 dw_2 \\
&= \frac{1}{\pi\sqrt{1-\rho^2}} \int_0^\infty \int_0^\infty \exp\{-\frac{w_2^1 - 2\rho w_1 w_2 + w_2^2}{2(1-\rho^2)}\} dw_1 dw_2 \\
&= \frac{1}{\pi\sqrt{1-\rho^2}} \int_0^{\frac{\pi}{2}} \int_0^\infty r \exp\{-\frac{r^2(1 - 2\rho\sin\theta\cos\theta)}{2(1-\rho^2)}\} dr d\theta \\
&= \frac{1}{\pi\sqrt{1-\rho^2}} \int_0^{\frac{\pi}{2}} \frac{1-\rho^2}{1-\rho\sin 2\theta} d\theta \\
&= \frac{\sqrt{1-\rho^2}}{\pi} \int_0^\pi \frac{1}{2(1-\rho\sin\theta)} d\theta \\
&= \frac{\sqrt{1-\rho^2}}{\pi} \frac{1}{\sqrt{1-\rho^2}} \arctan\left(\frac{\tan\frac{\theta}{2} - \rho}{\sqrt{1-\rho^2}}\right)\Big|_0^\pi \\
&= \frac{1}{\pi}\left[\frac{\pi}{2} + \arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)\right] \\
&= \frac{1}{2} + \frac{1}{\pi}\arcsin(\rho). \qquad (2.58)
\end{aligned}$$

By the linear transformation

$$\begin{pmatrix} Z_1^* \\ Z_2^* \end{pmatrix} = \begin{pmatrix} \frac{1}{2C_2} & -\frac{1}{2C_1} \\ \frac{1}{2C_2} & \frac{1}{2C_1} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \qquad (2.59)$$

which maps $R_1 = \{(x, y) : -C_1/C_2 x < y < C_1/C_2 x, x > 0\}$ into the first quadrant and $R_2 = \{(x, y) : C_1/C_2 x < y < -C_1/C_2 x, x < 0\}$ into the third quadrant, we convert calculating $P(C_1|Z_1| > C_2|Z_2|)$ into calculating $P(Z_1^* Z_2^* > 0)$. Under (2.59),

$$Z_1^* \sim \mathcal{N}(0, \frac{1}{4C_2^2} + \frac{1}{4C_1^2}), \qquad Z_2^* \sim \mathcal{N}(0, \frac{1}{4C_2^2} + \frac{1}{4C_1^2}),$$

and correlation of Z_1^* and Z_2^* is

$$\text{corr}(Z_1^*, Z_2^*) = \frac{C_1^2 - C_2^2}{C_1^2 + C_2^2}.$$

Finally, $P(C_1|Z_1| > C_2|Z_2|) = P(Z_1^* Z_2^* > 0) = \frac{1}{2} + \frac{1}{\pi} \arcsin\left(\frac{C_1^2 - C_2^2}{C_1^2 + C_2^2}\right)$. □

Now we can proceed and prove the theorem.

Proof of the Theorem: By Lemma 2.1, the difference in the loss function by introducing an extra variable into the model is a linear combination of two independent χ_1^2 random variables. Furthermore, by (2.53), (2.52) and 2.55),

$$\gamma_{s_1} \mathbf{Z}' \mathbf{x}^{(s_1)} \mathbf{x}^{(s_1)\prime} \mathbf{Z} \gamma_{s_2} \mathbf{Z}' \mathbf{x}^{(s_2)} \mathbf{x}^{(s_2)\prime} \mathbf{Z} = b^2 \gamma_{s_1} \gamma_{s_2} \mathbf{Z}' \mathbf{x}^{(s_1)} \tilde{\mathbf{x}}^{(s_1)\prime} \tilde{\mathbf{x}}^{(s_2)} \mathbf{x}^{(s_2)\prime} \mathbf{Z} = \mathbf{0},$$

and

$$\mathbf{x}(s_1) \tilde{\mathbf{x}}^{(s_1)\prime} \tilde{\mathbf{x}}^{(s_2)} \mathbf{x}^{(s_2)\prime} = 0.$$

Therefore, $nL_n(\alpha_{s_1}) - nL_n(\alpha_s)$ and $nL_n(\alpha_{s_2}) - nL_n(\alpha_s)$ are independent random variables. Let $\alpha_{s_1 s_2}$ be the model with both $\mathbf{x}^{s_1}$ and $\mathbf{x}(s_2)$ added into model α_s, then

$$\begin{aligned} nL_n(\alpha_{s_1 s_2}) - nL_n(\alpha_{s_1}) &= -\gamma_{s_2} \mathbf{u}' \mathbf{Z}' \mathbf{x}^{(s_2)} \mathbf{x}^{(s_2)\prime} \mathbf{Z} \mathbf{u} + \gamma_{s_2} \mathbf{e}' \mathbf{x}^{(s_2)} \mathbf{x}^{(s_2)\prime} \mathbf{e} \\ &= nL_n(\alpha_{s_2}) - nL_n(\alpha_s). \end{aligned} \tag{2.60}$$

In other words, the difference in loss function by adding one more variable does not depend on the variables that are already in the model. Therefore the order by which the variables are added into the model does not change the probability that this specific variable will reduce the loss function. Choosing the minimizer of the loss function is equivalent to a series of yes-no questions, starting from the smallest

model (with only the p^* true fixed-effects), and a "yes" means the loss function gets smaller by adding the variable. Therefore, the number of extra variables in the model follows a Binomial distribution, with parameters q_n, and P_a, which is the probability of answering "yes":

$$P_a = P[nL_n(\alpha_{s_k}) - nL_n(\alpha_s) < 0] = P[-b\sigma_u^2\eta_k + \sigma_e^2\zeta_k < 0] = P[\sqrt{b}\sigma_u|Z_1| > \sigma_e|Z_2|]$$

where Z_1 and Z_2 are independent standard normal variables. Using Lemma 2.2 with $C_1^2 = b\sigma_u^2$ and $C_2^2 = \sigma_e^2$,

$$P_a = \frac{1}{2} + \frac{1}{\pi}\arcsin\left(\frac{b\sigma_u^2 - \sigma_e^2}{b\sigma_u^2 + \sigma_e^2}\right).$$

□

Remark: The expected number of extra variables in this case is $q_n \cdot P_a$. The probability P_a is determined solely by the ratio $b\sigma_u^2/\sigma_e^2$. Moreover, $P_a \approx 0$ when $b\sigma_u^2 \ll \sigma_e^2$ and $P_a \approx 1$ when $\sigma_e^2 \ll b\sigma_u^2$. Therefore when $b\sigma_u^2$ is large enough compared to σ_e^2, the model that minimizes the loss function will choose a model that is very large, and the closer P_a is to 1, the bigger the final model is; on the other hand, if $b\sigma_u^2$ is relatively small (in which case, the model misspecification is negligible and is not an issue any more), the model that minimizes the loss function will only include the p^* fixed effects. □

Sequential Selection

The second special case we consider is sequential selection, when there is a specific order with which new variables are allowed into the model. In this case, if

the original design matrix is not orthogonal, we can first orthogonalize by Gram-Schmidt procedure. Since the i^{th} column of the orthogonalized matrix is a linear combination of the first i columns of the original matrix, selecting the i^{th} variable in the orthogonalized model means selecting the first i variables in the original matrix, so the selection order is preserved with the orthogonalization. Therefore, when there is an order, it does not matter if it is an orthogonal design matrix or not. But we do need balanced data structure to calculate some of the probabilities. Again we assume that only models that include all the p^* fixed effects are included in $\mathcal{A}_n$.

Assumption 2.4 *For any* $i \in \{1, \cdots, m\}$, $n_i = b$ *and* $\mathcal{A}_n = \{\alpha_{p^*+1}, \cdots, \alpha_{p_n}\}$, *where* $\alpha_k = \{1, \cdots, k\}$.

Theorem 2.8 *Under Assumption 2.4, the expected number of extra variables in the model that minimizes the loss function (2.51) is*

$$\frac{P_a(1 - P_a^{q_n})}{1 - P_a},$$

where $q_n = p_n - p^*$ *is the total number of "added" candidate variables and*

$$P_a = \frac{1}{2} + \frac{1}{\pi} arcsin\left(\frac{b\sigma_u^2 - \sigma_e^2}{b\sigma_u^2 + \sigma_e^2}\right).$$

Proof: The proof is exactly the same as Theorem 2.7, concerning the difference (2.57) and the distribution that this difference follows (Lemma 2.1). What is different is in the way we choose the final model: with a specific order, we stop asking "yes-no" questions once we get a "no"(adding the specific variable in the queue will not decrease the loss function), while without specific order, we don't stop until all

the q_n questions are asked. Therefore, if q_o is the number of extra variables in the final model of an order selection, with P_a being the probability of answering "yes",

$$P(q_o = 0) = 1 - P_a,$$

$$P(q_o = k) = P_a^k(1 - P_a)$$

for $k = 1, \cdots, q_n - 1$, and

$$P(q_o = q_n) = P_a^{q_n}$$

Therefore,

$$\begin{aligned}
Eq_o &= \sum_{k=0}^{q_n} kP(p_o = k) \\
&= (1 - P_a) \sum_{k=1}^{q_n - 1} kP_a^k + q_n P_a^{q_n} \\
&= \sum_{k=1}^{q_n} kP_a^k - \sum_{k=1}^{q_n} P_a^{k+1} + q_n P_a^{q_n} \\
&= \sum_{k=1}^{q_n - 1} P_a^k - (q_n - 1)P_a^{q_n} + q_n P_a^{q_n} \\
&= \frac{1 - P_a^{q_n+1}}{1 - P_a} - 1 \\
&= \frac{P_a(1 - P_a^{q_n})}{1 - P_a}
\end{aligned}$$

□

Note that again the number of extra variables chosen by a sequential selection will depend on the probability P_a. The expectation of q_o is close to 0 when $P_a \approx 0$, meaning that the final model chosen contains no extra variables; when $b\sigma_u^2$ is large compared to σ_e^2, on the other hand, since $(1 - P_a^{q_n})/(1 - P_a) \sim q_n$ when P_a is close to one, the expectation of q_o will be close to $P_a q_n$.

Remark: From the above two special cases, we see that when there is no correct model in $\mathcal{A}_n$, minimizing the loss function does not necessarily mean selecting a parsimonious model. In both of our cases, when $b\sigma_u^2$ is large compared to σ_e^2, the model minimizing the loss function will select a final model with approximately $q_n P_a$ extra variables. □

2.4.3 The Variance Estimators

There are many possible choices for the variance estimator in (2.50). We restrict our attention to

$$\hat{\sigma}_2^2(\alpha) = s^2(\alpha) = \frac{\mathbf{y}'(\mathbf{I}_n - \mathbf{H}_n(\alpha))\mathbf{y}}{n - p_n(\alpha)}. \tag{2.61}$$

This is the sample variance under the model α, and is a very popular choice of variance estimator. Sometimes people use n instead of $(n - p_n)$ in the denominator, but since $p_n/n \to 0$ as $n \to \infty$, this difference does not affect the limit or the distribution of the variance estimator, so we will only discuss the estimators defined in (2.61). For any $\alpha \in \mathcal{A}_n$ fixed, it is not hard to prove that $s^2(\alpha) \xrightarrow{p} \sigma_u^2 + \sigma_e^2$ following the same arguments as in Theorem 2.4. But since the size of $\mathcal{A}_n$ could be as large as 2^{p_n}, the uniform convergence over sets of models can not be taken granted and needs to be treated carefully.

Theorem 2.9 *Under Assumptions 1.2-1.6 and 2.1 the variance estimators* $s^2(\alpha) \xrightarrow{p} \sigma_u^2 + \sigma_e^2$ *uniformly for all* $\alpha \in \mathcal{A}_n$.

Proof: Since for any $\alpha \in \mathcal{A}_n$,

$$\begin{aligned} |s^2(\alpha) - (\sigma_u^2 + \sigma_e^2)| &= \left| \frac{1}{n - p_n(\alpha)} (\mathbf{Zu} + \mathbf{e})'(\mathbf{I}_n - \mathbf{H}_n(\alpha))(\mathbf{Zu} + \mathbf{e}) - (\sigma_u^2 + \sigma_e^2) \right| \\ &\leq \left| \frac{(\mathbf{Zu} + \mathbf{e})'(\mathbf{Zu} + \mathbf{e})}{n - p_n(\alpha)} - (\sigma_u^2 + \sigma_e^2) \right| \\ &+ \left| \frac{(\mathbf{Zu} + \mathbf{e})'\mathbf{H}_n(\alpha)(\mathbf{Zu} + \mathbf{e})}{n - p_n(\alpha)} \right| \end{aligned}$$

Let $e_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$ be the average error of the i^{th} cluster; then

$$\frac{(\mathbf{Zu} + \mathbf{e})'(\mathbf{Zu} + \mathbf{e})}{n - p_n(\alpha)} = \frac{\sum_{i=1}^m n_i u_i^2 + 2\sum_{i=1}^m n_i u_i e_{i\cdot} + \sum_{i=1}^m \sum_{j=1}^{n_i} e_{ij}^2}{n - p_n(\alpha)}.$$

Since the sequences u_i and e_{ij} are iid sequences, and $u_i e_{i\cdot}$ is a zero-mean independent sequence with finite fourth moment, and n_i's are assumed to be independent of u_i's and e_{ij}'s,

$$\frac{1}{n} \sum_i n_i u_i^2 = \frac{m}{n} \cdot \frac{1}{m} \sum_i n_i u_i^2 \xrightarrow{p} \sigma_u^2$$

by the Strong Law of Large Numbers and Slutsky's Theorem. Similarly,

$$\frac{1}{n} \sum_i n_i u_i e_{i\cdot} \xrightarrow{p} 0,$$

and

$$\frac{1}{n} \sum_i \sum_j e_{ij}^2 \xrightarrow{a.s.} \sigma_e^2.$$

Therefore it suffices to show that

$$\left| \frac{(\mathbf{Zu} + \mathbf{e})'\mathbf{H}_n(\alpha)(\mathbf{Zu} + \mathbf{e})}{n - p_n(\alpha)} \right| \xrightarrow{p} 0 \tag{2.62}$$

uniformly in $\alpha \in \mathcal{A}_n$. First of all, by Theorem 1.1,

$$(m^* - a_n)\mathbf{I}_{p_n} \leq m^{-1}(\mathbf{X}_n'\mathbf{X}_n) \leq (M^* + a_n)\mathbf{I}_{p_n}.$$

Then by Corollary 1.4

$$(M^* + a_n)^{-1} \leq \lambda_{\min}\left[\frac{1}{m}\left(\mathbf{X}_n'\mathbf{X}_n\right)^{-1}\right] \leq \lambda_{\max}\left[\frac{1}{m}\left(\mathbf{X}_n'\mathbf{X}_n\right)^{-1}\right] \leq (m^* - a_n)^{-1},$$

and

$$(M^* + a_n)^{-1}\mathbf{I}_{p_n} \leq m(\mathbf{X}_n'\mathbf{X}_n)^{-1} \leq (m^* - a_n)^{-1}\mathbf{I}_{p_n}.$$

The choosing matrix $\mathbf{I}_n(\alpha)$ has the properties

$$\mathbf{I}_n'(\alpha)\mathbf{I}_n(\alpha) = \mathbf{I}_{p_n}(\alpha),$$

and

$$\mathbf{I}_n'(\alpha)\mathbf{P}_1\mathbf{I}_n(\alpha) \leq \mathbf{I}_n'(\alpha)\mathbf{P}_2\mathbf{I}_n(\alpha)$$

if $\mathbf{P}_1 \leq \mathbf{P}_2$, and $\mathbf{I}_n'(\alpha)\mathbf{P}\mathbf{I}_n(\alpha)$ is nonnegative-definite if $\mathbf{P}$ is. Therefore

$$m^{-1}(M^* + a_n)^{-1}\mathbf{I}_{p_n(\alpha)} \leq (\mathbf{X}_n'(\alpha)\mathbf{X}_n(\alpha))^{-1} \leq m^{-1}(m^* - a_n)^{-1}\mathbf{I}_{p_n(\alpha)},$$

and

$$\mathbf{I}_n'(\alpha)\mathbf{X}_n'(\mathbf{Z}\mathbf{u} + \mathbf{e})'(\mathbf{Z}\mathbf{u} + \mathbf{e})\mathbf{X}_n\mathbf{I}_n(\alpha)$$

is nonnegative-definite. Moreover,

$$\begin{aligned}
& (\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{H}_n(\alpha)(\mathbf{Z}\mathbf{u} + \mathbf{e}) = |\mathrm{tr}[(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{H}_n(\alpha)(\mathbf{Z}\mathbf{u} + \mathbf{e})]| \\
= \;& \mathrm{tr}[(\mathbf{X}_n'(\alpha)\mathbf{X}_n(\alpha))^{-1}\mathbf{I}_n'(\alpha)\mathbf{X}_n'(\mathbf{Z}\mathbf{u} + \mathbf{e})(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}_n\mathbf{I}_n(\alpha)] \\
\leq \;& m^{-1}(m^* - a_n)^{-1}\mathrm{tr}[\mathbf{I}_n'(\alpha)\mathbf{X}_n'(\mathbf{Z}\mathbf{u} + \mathbf{e})(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}_n\mathbf{I}_n(\alpha)] \\
\leq \;& \frac{p_n(\alpha)}{m(m^* - a_n)}\lambda_{\max}\left(\mathbf{I}_n'(\alpha)\mathbf{X}_n'(\mathbf{Z}\mathbf{u} + \mathbf{e})(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}_n\mathbf{I}_n(\alpha)\right) \\
= \;& \frac{p_n(\alpha)}{m(m^* - a_n)}\left\|\mathbf{I}_n'(\alpha)\mathbf{X}_n'(\mathbf{Z}\mathbf{u} + \mathbf{e})(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}_n\mathbf{I}_n(\alpha)\right\| \\
\leq \;& \frac{p_n(\alpha)}{m(m^* - a_n)}\left\|(\mathbf{Z}\mathbf{u} + \mathbf{e})'\mathbf{X}_n\mathbf{I}_n(\alpha)\right\|^2
\end{aligned}$$

$$\begin{aligned} &\leq \frac{p_n(\alpha)}{m(m^*-a_n)} p_n(\alpha) \max_{k\in\alpha} |(\mathbf{X}_n'(\mathbf{Zu}+\mathbf{e}))_k| \\ &\leq \frac{p_n^2}{m(m^*-a_n)} \max_{1\leq k\leq p_n} \left|\sum_{i=1}^m \tilde{\mathbf{x}}_{ik} n_i (u_i + e_{i\cdot})\right| \end{aligned} \tag{2.63}$$

The sequence $\{\tilde{\mathbf{x}}_i n_i (u_i + e_{i\cdot})\}_i$ is an iid sequence with 0-mean and $4r^{th}$ moment. By arguments similar to Theorem 1.1,

$$\begin{aligned} & P\left[\max_{1\leq k\leq p_n} \left|\frac{1}{m}\sum_{i=1}^m \mathbf{x}_{ik} n_i (u_i + e_{i\cdot})\right| > \frac{(n-p^*)(m^*-a_n)}{p_n^2}\right] \\ \leq\ & p_n \max_{1\leq k\leq p_n} P\left[\left|\frac{1}{m}\sum_{i=1}^m \mathbf{x}_{ik} n_i (u_i + e_{i\cdot})\right| > \frac{(n-p^*)(m^*-a_n)}{p_n^2}\right] \\ \leq\ & M_r p_n^{8r+1} n^{-2r} (n-p^*)^{-4r} \\ =\ & O(n^{(8r+1)\theta - 6r}) \end{aligned} \tag{2.64}$$

where M_r is a constant over n. Finally, if $r > 2\theta/(1-2\theta) > \theta/(6-8\theta)$, for any $0 < \epsilon < (2r(3-4\theta)-\theta)/4r$,

$$\begin{aligned} & P\left[\frac{1}{n-p_n(\alpha)}(\mathbf{Zu}+\mathbf{e})'\mathbf{H}_n(\alpha)(\mathbf{Zu}+\mathbf{e}) \geq n^{-\epsilon}\right] \\ \leq\ & P\left[\max_{1\leq k\leq p_n} \left|\frac{1}{m}\sum_{i=1}^m \mathbf{x}_{ik} n_i (u_i + e_{i\cdot})\right| \geq \frac{(n-p^*)(m^*-a_n)}{p_n^2 n^{\epsilon}}\right] \\ =\ & O(n^{(8r+1)\theta - 6r + 4r\epsilon}) \to 0 \end{aligned} \tag{2.65}$$

Therefore (2.62) is proved. □

The following theorem provides the bound for $E|s^2 - (\sigma_u^2 + \sigma_e^2)|$:

Theorem 2.10 *Under Assumptions 1.2-1.4 and 2.1,*

$$E|s^2 - (\sigma_u^2 + \sigma_e^2)| = O(n^{-1/2}).$$

Proof: First of all,

$$E|s^2 - (\sigma_u^2 + \sigma_e^2)| = E|(n-p_n)^{-1}(\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n - \mathbf{H}_n)(\mathbf{Zu}+\mathbf{e}) - (\sigma_u^2 + \sigma_e^2)|$$

$$\leq \quad E|(n-p_n)^{-1}(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})-(\sigma_u^2+\sigma_e^2)|$$

$$-E|(n-p_n)^{-1}(\mathbf{Zu}+\mathbf{e})'\mathbf{H}_n(\mathbf{Zu}+\mathbf{e})|. \tag{2.66}$$

The random vector $(\mathbf{Zu}+\mathbf{e})$ is a $n\times 1$ vector of zero-mean random variables with variance-covariance matrix $\mathbf{V}=\sigma_u^2\mathbf{Z}'\mathbf{Z}+\sigma_e^2\mathbf{I}_n$. The quadratic form $(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})$ therefore has moments

$$E[(\mathbf{Zu}+\mathbf{e})'\mathbf{Zu}+\mathbf{e})]=\mathrm{tr}[\mathbf{V}]=n(\sigma_u^2+\sigma_e^2),$$

and

$$E[(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})-n(\sigma_u^2+\sigma_e^2)]^2=2\mathrm{tr}[\mathbf{V}^2]=2\left[\sigma_u^4\sum_{i=1}^m n_i^2+n\sigma_e^4+2n\sigma_u^2\sigma_e^2\right]=O(n).$$

Therefore

$$\begin{aligned}
& E|(n-p_n)^{-1}(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})-(\sigma_u^2+\sigma_e^2)| \\
\leq\ & (n-p_n)^{-1}E|(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})-n(\sigma_u^2+\sigma_e^2)|+\left(\frac{n}{n-p_n}-1\right)(\sigma_u^2+\sigma_e^2) \\
\leq\ & (n-p_n)^{-1}\sqrt{E[(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})-n(\sigma_u^2+\sigma_e^2)]^2}+O(p_n/n) \\
=\ & O(n^{-1/2})+O(p_n/n).
\end{aligned} \tag{2.67}$$

If we let $\eta_i^{(k)}\equiv\tilde{\mathbf{x}}_{ik}n_i(u_i+e_{i\cdot})$, then $\eta_i^{(k)}$ are iid zero-mean random variables with finite $(4r)^{th}$ moment uniformly in k. Moreover,

$$E\left[\sum_{i=1}^m\tilde{\mathbf{x}}_{ik}n_i(u_i+e_{i\cdot})\right]^2=E\left[\sum_i\eta_i^{(k)}\right]^2\overset{iid}{=}\sum_{i=1}^m E\left[\eta_i^{(k)}\right]^2\leq mC^* \tag{2.68}$$

for all $1\leq k\leq p_n$. Note that

$$m^{-1}(M^*+a_m)^{-1}\mathbf{I}_{p_n}\leq(\mathbf{X}'\mathbf{X})^{-1}\leq m^{-1}(m^*-a_m)^{-1}\mathbf{I}_{p_n}$$

and

$$\mathbf{X}'(\mathbf{Zu}+\mathbf{e})'(\mathbf{Zu}+\mathbf{e})\mathbf{X} = [\mathbf{X}'(\mathbf{Zu}+\mathbf{e})]^{\otimes 2}$$

is positive definite By Corollary A.1 and Definition 1.7,

$$\begin{aligned}
& \operatorname{tr}\left[(\mathbf{Au}+\mathbf{e})'\mathbf{H}_n(\mathbf{Zu}+\mathbf{e})\right] = \operatorname{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Zu}+\mathbf{e})(\mathbf{Zu}+\mathbf{e})'\mathbf{X}\right] \\
\overset{Col.A.1}{\leq} \quad & m^{-1}(m^*-a_m)^{-1}\operatorname{tr}\left[\mathbf{X}'(\mathbf{Zu}+\mathbf{e})(\mathbf{Zu}+\mathbf{e})'\mathbf{X}\right] \\
\leq \quad & m^{-1}(m^*-a_m)^{-1}p_n\lambda_{\max}\left[\mathbf{X}'(\mathbf{Zu}+\mathbf{e})(\mathbf{Zu}+\mathbf{e})'\mathbf{X}\right] \\
\overset{Def.1.7}{=} \quad & m^{-1}(m^*-a_m)^{-1}p_n\left\|\mathbf{X}'(\mathbf{Zu}+\mathbf{e})(\mathbf{Zu}+\mathbf{e})'\mathbf{X}\right\| \\
\leq \quad & p_n m^{-1}(m^*-a_m)^{-1}\left\|\mathbf{X}'(\mathbf{Zu}+\mathbf{e})\right\|^2 \\
= \quad & p_n m^{-1}(m^*-a_m)^{-1}\sum_{k=1}^{p_n}\left[\sum_{i=1}^{m}\tilde{\mathbf{x}}_{ik}n_i(u_i+e_{i\cdot})\right]^2. \qquad (2.69)
\end{aligned}$$

Therefore

$$\begin{aligned}
& (n-p_n)^{-1}E|(\mathbf{Zu}+\mathbf{e}'\mathbf{H}_n(\mathbf{Zu}+\mathbf{e})| \\
\leq \quad & \frac{p_n(n-p_n)^{-1}}{m(m^*-a_m)}\sum_{k=1}^{p_n}E\left[\sum_{i=1}^{m}\tilde{\mathbf{x}}_{ik}n_i(u_i+e_{i\cdot})\right]^2 \\
\overset{(2.68)}{\leq} \quad & \frac{p_n(n-p_n)^{-1}}{m(m^*-a_m)}\cdot p_n\cdot mC^* \\
= \quad & O(p_n^2/n) \qquad (2.70)
\end{aligned}$$

In conclusion,

$$E|s^2-(\sigma_u^2+\sigma_e^2)| = O(n^{-1/2}) + O(p_n/n) + O(p_n^2/n) = O(n^{-1/2})$$

when $\theta < 1/4$. □

Remark: Not only does s^2 converge to $(\sigma_u^2+\sigma_e^2)$ in probability, but the distance $|s^2-(\sigma_u^2+\sigma_e^2)|$ is of order $(n^{-1/2+\varepsilon})$ in probability and of order $n^{-1/2})$ in L^1 norm, namely, the distance $|s^2-(\sigma_u^2+\sigma_e^2)|$ is integrable. □

Since s^2 converges to a point $(\sigma_u^2 + \sigma_e^2)$, the expectation of certain functions should be well approximated by the function evaluated at the point $(\sigma_u^2 + \sigma_e^2)$. The following corollary states this result:

Corollary 2.3 *Let $g_n(x)$ be a sequence of piecewise smooth functions defined on $[0, \infty)$. Suppose that there exist N_g and $0 < C_g < (\sigma_u^2 + \sigma_e^2)$ such that $g_n(x)$ are uniformly bounded for $0 \leq x \leq C_g$ and $g_n'(x)$ exist and are uniformly bounded for $x \geq C_g$ for all $n \geq N_g$. Let the nonrandom sequence $t_n > 2$ be $O(n^{1/2-\delta})$ with $0 < \delta \leq 1/2$. Then*

$$|E[g_n(t_n s^2)] - g(t_n(\sigma_u^2 + \sigma_e^2))| = O(n^{-\delta}).$$

Proof: First of all, with $0 < C_g < \sigma_u^2 + \sigma_e^2$ and $2 < t_n = O(n^{1/2-\delta})$, $C_g/t_n < \sigma_u^2 + \sigma_e^2$ and $|C_g/t_n - (\sigma_u^2 + \sigma_e^2)| = (\sigma_u^2 + \sigma_e^2) + O(t_n^{-1})$. Therefore by Chebyshev's inequality

$$\begin{aligned} P[s^2 \leq C_g/t_n] &\leq P[|s^2 - (\sigma_u^2 + \sigma_e^2)| \geq (\sigma_u^2 + \sigma_e^2) + O(t_n^{-1})] \\ &\leq O(n^{-1}) \left[(\sigma_u^2 + \sigma_e^2) + O(t_n^{-1})\right]^2 \\ &= O(n^{-1}) \end{aligned} \tag{2.71}$$

since $O(t_n^{-1}) \leq O(1)$. Let M_g be the uniform bound of $|g_n|$ on $[0, C_g]$ and $M_{g'}$ be the uniform bound of $|g_n'|$ on $[C_g, \infty)$. Then

$$E\left[|g(t_n s^2) - g(t_n(\sigma_u^2 + \sigma_e^2))| I_{[t_n s^2 \leq C_g]}\right] \leq 2M_g P[t_n s^2 \leq C_g] = O(n^{-1})$$

and

$$E\left[|g(t_n s^2) - g(t_n(\sigma_u^2 + \sigma_e^2))| I_{[t_n s^2 \geq C_g]}\right] \leq E\left[M_{g'} t_n |s^2 - (\sigma_u^2 + \sigma_e^2)| I_{[t_n s^2 \geq C_g]}\right]$$

$$\leq M_{g'} t_n E[|s^2 - (\sigma_u^2 + \sigma_e^2)|]$$

$$\leq O(n^{1/2-\delta}) O(n^{-1/2}) = O(n^{-\delta}).$$

Therefore

$$\begin{aligned} & |E[g_n(t_n s^2)] - g_n(t_n(\sigma_u^2 + \sigma_e^2))| \\ \leq \; & E\left[|g_n(t_n s^2) - g_n(t_n(\sigma_u^2 + \sigma_e^2))|\right] \\ \leq \; & O(n^{-1}) + O(n^{-\delta}) \\ = \; & O(n^{-\delta}). \end{aligned}$$

□

Remark: From the proof of Corollary 2.3 we can see that the result still holds if we let C_g depend on n but be uniformly bounded away from $(\sigma_u^2 + \sigma_e^2)$.

2.4.4 The Minimizer of Γ_{n,λ_n}

We mainly discuss two types of variance estimator in (2.50):

$$\Gamma_{n,\lambda_n}^{G_1}(\alpha) = n^{-1}\|\mathbf{y} - \hat{\mathbf{y}}(\alpha)\|^2 + n^{-1}\lambda_n p_n(\alpha) s^2 \tag{2.72}$$

where s^2 is the sample variance of the full model, and

$$\Gamma_{n,\lambda_n}^{G_2}(\alpha) = n^{-1}\|\mathbf{y} - \hat{\mathbf{y}}(\alpha)\|^2 + n^{-1}\lambda_n p_n(\alpha) s^2(\alpha) \tag{2.73}$$

where $s^2(\alpha)$ is defined in (2.61). Thus, $\Gamma_{n,\lambda_n}^{G_1}$ has a universal variance estimator for all the $\alpha \in \mathcal{A}_n$ and $\Gamma_{n,\lambda_n}^{G_2}$ has different variance estimators for each α.

Analytical results about the final model chosen by $\Gamma_{n,\lambda_n}^{G_1}$ can be obtained only in some special cases, but analytical results about the final model chosen by $\Gamma_{n,\lambda_n}^{G_2}$ are not available even in the simplest special cases.

Balanced Data, Orthogonal Design

Theorem 2.11 *Under Assumption 2.3, with* $\lambda_n = o(n^{\frac{1}{2}-\delta})$ *for* $0 < \delta \leq 1/2$, *then the model that minimizes* $\Gamma^{G_1}_{n,\lambda_n}$ *contains* $(p^* + p_1)$ *variables, where given* s^2, p_1 *follows a Binomial distribution. The unconditional expectation is* $Ep_1 = q_n P_a^{G_1}$, *with*

$$P_a^{G_1} = 2\left(1 - \Phi\left(\sqrt{\frac{\lambda_n(\sigma_e^2 + \sigma_u^2)}{b\sigma_u^2 + \sigma_e^2}}\right)\right) + O(n^{-\delta}). \tag{2.74}$$

Proof: With $s^2 = \hat{\sigma}^2(\alpha)$ as in (2.50), then

$$n\Gamma^{G_1}_{n,\lambda_n}(\alpha) = (\mathbf{Zu} + \mathbf{e})'(\mathbf{I}_n - \mathbf{H}_n(\alpha))(\mathbf{Zu} + \mathbf{e}) + \lambda_n p_n(\alpha) s^2;$$

and for two models α_s and α_{s_1} where $\mathbf{X}(\alpha_{s_1}) = (\mathbf{X}(\alpha_s)|\mathbf{x}^{(s_1)})$, the difference in their GIC_{λ_n} is

$$n\Gamma^{G_1}_{n,\lambda_n}(\alpha_{s_1}) - n\Gamma^{G_1}_{n,\lambda_n}(\alpha_s) = -\gamma_{s_1}(\mathbf{Zu} + \mathbf{e})'\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}(\mathbf{Zu} + \mathbf{e}) + \lambda_n s^2. \tag{2.75}$$

For the normal vector $(\mathbf{Zu} + \mathbf{e}) \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ where $\mathbf{V} = \sigma_e^2\mathbf{I}_n + \sigma_u^2\mathbf{ZZ}'$,

$$\begin{aligned}
\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}\mathbf{V} &= \sigma_e^2\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime} + \sigma_u^2\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}\mathbf{ZZ}' \\
&\overset{(2.53)}{=} \sigma_e^2\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime} + \sigma_u^2\gamma_{s_1}\mathbf{x}^{(s_1)}\tilde{\mathbf{x}}^{(s_1)\prime}\mathbf{Z}'\mathbf{ZZ}' \\
&\overset{(2.52)}{=} \sigma_e^2\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime} + b\sigma_u^2\gamma_{s_1}\mathbf{x}^{(s_1)}\tilde{\mathbf{x}}^{(s_1)\prime}\mathbf{Z}' \\
&\overset{(2.53)}{=} (\sigma_e^2 + b\sigma_u^2)\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}.
\end{aligned}$$

Since $(\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime})$ is idempotent, the rank-1 matrix $[(\sigma_e^2 + b\sigma_u^2)^{-1}\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}\mathbf{V}]$ is also idempotent, and therefore the quadratic form

$$(b\sigma_u^2 + \sigma_e^2)^{-1}(\mathbf{Zu} + \mathbf{e})'\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}\mathbf{Zu} + \mathbf{e}) = \eta_{s_1} \sim \chi^2_1,$$

and the probability that the difference (2.75) is less than 0 is

$$\begin{aligned} P_a^{G_1}(s^2) &\equiv P[-\gamma_{s_1}(\mathbf{Zu}+\mathbf{e})'\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}(\mathbf{Zu}+\mathbf{e})+\lambda_n s^2<0] \\ &= P[-(b\sigma_u^2+\sigma_e^2)\eta_{s_1}+\lambda_n s^2<0] \\ &= P\left[\eta_{s_1}>\frac{\lambda_n s^2}{b\sigma_u^2+\sigma_e^2}\right] \end{aligned} \tag{2.76}$$

Note that again

$$n\Gamma_{n,\lambda_n}(\alpha_{s_1 s_2})-n\Gamma_{n,\lambda_n}(\alpha_{s_1})=n\Gamma_{n,\lambda_n}(\alpha_{s_2})-n\Gamma_{n,\lambda_n}(\alpha_s). \tag{2.77}$$

This means that the difference is the same regardless of the other terms in the model. Again selecting the model that minimizes the loss function involves a series of "yes-no" questions, with "yes" meaning that adding the variable will reduce the loss function. The difference in loss function by adding the k^{th} variable into the model is

$$-\gamma_k(\mathbf{Zu}+\mathbf{e})'\mathbf{x}^{(k)}\mathbf{x}^{(k)\prime}(\mathbf{Zu}+\mathbf{e})+\lambda_n s^2.$$

All the differences share the same variable s^2, so the series of questions can not be independent. Note that

$$s^2=(n-p_n)^{-1}(\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n-\mathbf{H}_n)(\mathbf{Zu}+\mathbf{e})$$

is also a quadratic form of normals. From Searle [23] Chapter 2, we know that two quadratic from of normal variables $\mathbf{v}'\mathbf{M}_1\mathbf{v}$ and $\mathbf{v}'\mathbf{M}_2\mathbf{v}$ for $\mathbf{v}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}_v)$ are independent if and only if $\mathbf{M}_1\boldsymbol{\Sigma}_v\mathbf{M}_2=\mathbf{0}$. Since

$$\begin{aligned} &\mathbf{x}^{(k)}\mathbf{x}^{(k)\prime}(\sigma_u^2\mathbf{ZZ}'+\sigma_e^2\mathbf{I}_n)(\mathbf{I}_n-\mathbf{H}_n) \\ =\ &\sigma_u^2(\mathbf{x}^{(k)}\mathbf{x}^{(k)\prime}(\mathbf{ZZ}'-\mathbf{ZZ}'\mathbf{H}_n))+\sigma_e^2\mathbf{x}^{(k)}\mathbf{x}^{(k)\prime}(\mathbf{I}_n-\mathbf{H}_n) \end{aligned}$$

$$
\begin{aligned}
&= \sigma_u^2(\mathbf{x}^{(k)}\tilde{\mathbf{x}}^{(k)}\mathbf{Z}'\mathbf{Z}\mathbf{Z}' - \mathbf{x}^{(k)}\mathbf{x}^{(k)'}\mathbf{Z}\mathbf{Z}'\mathbf{Z}\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}) \\
&= b\sigma_u^2\mathbf{x}^{(k)}\mathbf{x}^{(k)'}(\mathbf{I}_n - \mathbf{H}_n) \\
&= 0, \qquad (2.78)
\end{aligned}
$$

the two random variables η_k and s^2 are independent for $k = p^* + 1, \cdots, q_n$.

To select the optimal model that minimizes $\Gamma_{n,\lambda_n}^{G_1}$,

- There are altogether q_n "yes-no" questions asked. Whether or not a variable should be included in the chosen model depend on the answer to the question: does adding this variable reduce $\Gamma_{n,\lambda_n}^{G_1}$? The order in which we ask these questions does not matter, by (2.77);

- The probability of answering "yes" to the question is the same for each variable, by (2.76);

- The differences in $\Gamma_{n,\lambda}^{G_1}$ between adding one variable and adding another are not independent. But each difference is a linear combination of two independent variables, and given a common variable that is shared by all the differences, they are independent.

Therefore, we can calculate the expected number of p_1 by conditioning on the common variable s^2 first. Given s^2, the differences $[-(b\sigma_u^2+\sigma_e^2)\eta_k+\lambda_n s^2]$ for $k \in \{p^*+1, \cdots, p_n\}$ are independent variables with the same distribution. Therefore, given s^2 the number of extra variables p_1 follows a Binomial distribution with parameters q_n and $P_a^{G_1}(s^2)$. Hence

$$
Ep_1 = E[E[p_1|s^2]] = 2q_nE\left[1 - \Phi\left(\sqrt{\frac{\lambda_n s^2}{b\sigma_u^2 + \sigma_e^2}}\right)\right]
$$

Let $g_n(x) = 2(1-\Phi(\sqrt{x}))$ and $t_n = \lambda_n/(b\sigma_u^2+\sigma_e^2)$ in Corollary 2.3. Then $g_n(x)$ is uniformly bounded for all n and x, and since $g_n'(x) = -x^{-1/2}\phi(\sqrt{x})$, $g_n'(x)$ is uniformly bounded for $x \geq (\sigma_u^2+\sigma_e^2)/2$. Therefore by Corollary 2.3,

$$Ep_1 = q_n E[g_n(\lambda_n s^2)] = q_n\left[g_n(\lambda_n(\sigma_u^2+\sigma_e^2)) + o(n^{-\delta})\right] = q_n P_a^{G_1}$$

where $P_a^{G_1} = 2\left[1-\Phi\left(\sqrt{\frac{\lambda_n(\sigma_u^2+\sigma_e^2)}{b\sigma_u^2+\sigma_e^2}}\right)\right] + O(n^{-\delta})$. □

When we use (2.61) as the variance estimator in (2.50), for two models α_s and α_{s_1} where $\mathbf{X}(\alpha_{s_1}) = (\mathbf{X}(\alpha)|\mathbf{x}^{(s_1)})$ and $p_s = p_n(\alpha_s)$,

$$\begin{aligned} n\Gamma_{n,\lambda_n}^{G_2}(\alpha_s) &= \|\mathbf{y}-\hat{\mathbf{y}}(\alpha_s)\|^2 + \lambda_n p_s s^2(\alpha_s) \\ &= \mathbf{y}'(\mathbf{I}_n - \mathbf{H}_n(\alpha_s))\mathbf{y} + \frac{\lambda_n p_s}{n-p_s}\mathbf{y}'(\mathbf{I}_n-\mathbf{H}_n(\alpha_s))\mathbf{y} \\ &= \left(1+\frac{\lambda_n p_s}{n-p_s}\right)(\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n-\mathbf{H}_n(\alpha_s))(\mathbf{Zu}+\mathbf{e}) \end{aligned}$$

and

$$\begin{aligned} n\Gamma_{n,\lambda_n}^{G_2}(\alpha_{s_1}) &= \left(1+\frac{\lambda_n(p_s+1)}{n-p_s-1}\right)(\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n-\mathbf{H}_n(\alpha_{s_1}))(\mathbf{Zu}+\mathbf{e}) \\ &= \left(1+\frac{\lambda_n(p_s+1)}{n-p_s-1}\right)(\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n-\mathbf{H}_n(\alpha_s)-\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime})(\mathbf{Zu}+\mathbf{e}). \end{aligned}$$

Therefore,

$$\begin{aligned} & n\Gamma_{n,\lambda_n}^{G_2}(\alpha_{s_1}) - n\Gamma_{n,\lambda_n}^{G_2}(\alpha_s) \\ &= \lambda_n\left[\frac{p_s+1}{n-p_s-1}-\frac{p_s}{n-p_s}\right](\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n-\mathbf{H}_n(\alpha_s)(\mathbf{Zu}+\mathbf{e}) \\ &\quad -\left(1+\lambda_n\frac{p_s+1}{n-p_s-1}\right)(\mathbf{Zu}+\mathbf{e})'\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}(\mathbf{Zu}+\mathbf{e}) \\ &= \frac{n\lambda_n}{(n-p_s)(n-p_s-1)}(\mathbf{Zu}+\mathbf{e})'(\mathbf{I}_n-\mathbf{H}_n(\alpha_s))(\mathbf{Zu}+\mathbf{e}) \\ &\quad -\left(1+\lambda_n\frac{p_s+1}{n-p_s-1}\right)(\mathbf{Zu}+\mathbf{e})'\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}(\mathbf{Zu}+\mathbf{e}) \\ &= \frac{n\lambda_n}{n-p_s-1}s^2(\alpha_s) - \left(1+\lambda_n\frac{p_s+1}{n-p_s-1}\right)(b\sigma_u^2+\sigma_e^2)\eta_{s_1} \end{aligned}$$

where

$$\eta_{s_1} \equiv (b\sigma_u^2 + \sigma_e^2)^{-1}(\mathbf{Zu}+\mathbf{e})'\gamma_{s_1}\mathbf{x}^{(s_1)}\mathbf{x}^{(s_1)\prime}(\mathbf{Zu}+\mathbf{e}) \sim \chi_1^2.$$

Evidently,

$$n\Gamma_{n,\lambda_n}^{G_2}(\alpha_{s_1 s_2}) - n\Gamma_n n, \lambda^{G_2}(\alpha_{s_1}) \neq n\Gamma_{n,\lambda_n}^{G_2}(\alpha_{s_2}) - n\Gamma_{n,\lambda_n}^{G_2}(\alpha_s).$$

So the probability that adding $\mathbf{x}^{(s_2)}$ reduces GIC_{λ_n} depends on the variables that are already in the model. Finding the model that minimizes (2.50) is no longer as simple as asking "yes-no" questions, and the approximate distribution of the number of extra variables is not available.

Sequential Selection

Theorem 2.12 *Under Assumption 2.4, with* $\lambda_n = O(n^{\frac{1}{2}-\delta})$ *for* $0 < \delta \leq 1/2$, *the expected number of extra variables* $E[p_o^{G_1}]$ *in the model that minimizes* $\Gamma_{n,\lambda_n}^{G_1}$ *is*

$$E[p_o^{G_1}] = \frac{P_{G_1}(1-[P_{G_1}]^{q_n})}{1-P_{G_1}} + O(n^{-\delta})$$

where $P_{G_1} = 2\left[1-\Phi\left(\sqrt{\frac{\lambda_n(\sigma_u^2+\sigma_e^2)}{b\sigma_u^2+\sigma_e^2}}\right)\right]$.

Proof: In sequential selection, the difference in GIC_{λ_n} of two subsequent models is

$$n\Gamma_{n,\lambda_n}^{G_1}(\alpha_k) - n\Gamma_{n,\lambda_n}^{G_1}(\alpha_{k-1}) = -(\mathbf{Zu}+\mathbf{e})'\gamma_k\mathbf{x}^{(k)}\mathbf{x}^{(k)\prime}(\mathbf{Zu}+\mathbf{e}) + \lambda_n s^2.$$

As we proved in Theorem 2.11, this difference is a linear combination of two independent quadratic forms of normal variables, and

$$P[n\Gamma_{n,\lambda_n}^{G_1}(\alpha_k) - n\Gamma_{n,\lambda_n}^{G_1}(\alpha_{k-1}) < 0] = P[\eta_k > \frac{\lambda_n s^2}{b\sigma_u^2+\sigma_e^2}].$$

Conditional on s^2, the differences in $\Gamma^{G_1}_{n,\lambda_n}$ by adding one variable are independent, and the let expected number of extra variables be $p_o^{G_1}$, then

$$P[p_o^{G_1} = k|s^2] = [P_{G_1}(s^2)]^k(1 - P_{G_1}(s^2))$$

for $1 \leq k \leq q_n - 1$, and

$$P[p_o^{G_1} = q_n|s^2] = [P_{G_1}(s^2)]^{q_n}$$

where

$$P_{G_1}(s^2) \equiv 2\left(1 - \Phi\left(\sqrt{\frac{\lambda_n s^2}{b\sigma_u^2 + \sigma_e^2}}\right)\right).$$

The conditional expectation of $p_o^{G_1}$ is therefore

$$E[p_0^{G_1}|s^2] = P_{G_1}(s^2)\frac{1 - [P_{G_1}(s^2)]^{q_n}}{1 - P_{G_1}(s^2)}.$$

Let $h_n(x) = x(1 - x^{q_n})/(1 - x)$. In the interval $[0, \delta_*]$ for any $\delta_* < 1$, the functions $h_n(x)$ satisfy

$$h_n(x) = \frac{x(1 - x^{q_n})}{1 - x} \leq \frac{x}{1 - x} \leq \frac{\delta_*}{1 - \delta_*}$$

and

$$h_n'(x) = \frac{(1 - x^{q_n}) - q_n x^{q_n}(1 - x)}{(1 - x)^2} \leq \frac{1}{(1 - x)^2} \leq \frac{1}{(1 - \delta_*)^2}.$$

Let $f(x) = 2(1 - \Phi(\sqrt{x}))$ and $\delta_* = f(\epsilon_*)$. Then $f(x) \in [0, \delta_*]$ when $x \in [\epsilon_*, \infty)$. Finally let $g_n(x) = h(f(x))$ and $t_n = \lambda_n/(b\sigma_u^2 + \sigma_e^2)$. We can see that on the interval $[\epsilon_*, \infty)$, the functions $g_n(x)$ are uniformly bounded by $\delta_*/(1 - \delta_*)$ and $g_n'(x)$ are uniformly bounded by $1/(1 - \delta_*)^2 \cdot \epsilon_*^{-1/2}\phi(\sqrt{\epsilon_*})$. Therefore we can apply Corollary 2.3 with $C_g = \epsilon_*$ and conclude that

$$E[p_o^{G_1} I_{[t_n s^2 \geq \epsilon_*]}] = \frac{P_{G_1}(1 - [P_{G_1}]^{q_n})}{1 - P_{G_1}} I_{[t_n(\sigma_u^2 + \sigma_e^2) \geq \epsilon_*]} + O(n^{-\delta}).$$

Since t_n is bounded below,

$$E[p_o^{G_1} I_{[t_n s^2 \geq \delta_*]}] = \frac{P_{G_1}(1 - [P_{G_1}]^{q_n})}{1 - P_{G_1}} + O(n^{-\delta})$$

for ϵ_* sufficiently small. Now on the interval $[\delta_*, 1]$ the functions $h_n(x)$ are bounded by q_n. By Chebyshev's inequality and the fact that $\text{var}(s^2) = O(n^{-1})$,

$$P[t_n s^2 \leq \epsilon_*] \leq P[|s^2 - (\sigma_u^2 + \sigma_e^2)| \geq (\sigma_u^2 + \sigma_e^2) - \epsilon_*/t_n] \leq O(n^{-1})$$

since t_n is bounded below. Therefore

$$E[p_o^{G_1} I_{[t_n s^2 \geq \epsilon_*]}] \leq q_n P[t_n s^2 \leq \epsilon_*] = O(p_n n^{-1}).$$

The theorem is proved by noting that $p_n = O(n^\theta)$ with $\theta < 1/4$. □

2.4.5 The Various Selection Criteria

As Shao mentioned in his paper, GIC_{λ_n} unifies many other model selection methods. We say two selection methods are equivalent if asymptotically the functions they minimize are equal. In this section we discuss the equivalence of GIC_{λ_n} to other popular selection methods and the number of extra variables in the final model chosen by these various selection methods.

Let

$$S_n(\alpha) = \mathbf{y}'(\mathbf{I}_n - \mathbf{H}_n(\alpha))\mathbf{y} = (\mathbf{Zu} + \mathbf{e})'(\mathbf{I}_n - \mathbf{H}_n(\alpha))(\mathbf{Zu} + \mathbf{e})$$

be the residual sum of squares for model α, and $s^2(\alpha)$ and s^2 are defined in (2.61) and (2.23), respectively. Table 2.1 demonstrates the relationship between GIC_{λ_n} and these selection procedures.

Table 2.1: Comparison of GIC_{λ_n} to other Selection Methods

Name	Criterion	$\hat{\sigma}_n^2(\alpha)$	λ_n	Reference
AIC	$n\log(n^{-1}S_n(\alpha)) + 2p_n(\alpha)$	$s^2(\alpha)$	2	Akaike
BIC	$n\log(n^{-1}S_n(\alpha)) + (\log n)p_n(\alpha)$	$s^2(\alpha)$	$\log n$	Schwartz
C_p	$S_n(\alpha) + 2p_n(\alpha)s^2$	s^2	2	Mallows
FPE_λ	$S_n(\alpha) + \lambda p_n(\alpha)s^2$	s^2	$\lambda > 2$	Shibata [25]
GIC	$S_n(\alpha) + C_n p_n(\alpha)s^2$	s^2	C_n	Rao and Wu[22]

The sequence C_n in the table satisfies

$$C_n \to \infty, \quad \frac{C_n}{n} \to 0, \quad \frac{C_n}{\log\log n} \to \infty.$$

Note that Mallows' C_p, Shibata's FPE_λ and Rao and Wu's GIC are special cases of GIC_{λ_n}; the criteria AIC and BIC are equivalent to GIC_{λ_n} in the sense that asymptotically, the minimizer of AIC in $\mathcal{A}_n$ is the same as the minimizer of GIC_{λ_n} and the minimizer of BIC is the same as that of GIC_{λ_n} in $\mathcal{A}_n$. To see this equivalence, note that when we use $s^2(\alpha)$ as $\hat{\sigma}^2(\alpha)$ in (2.50),

$$\Gamma_{n,\lambda_n}^{G_2}(\alpha) = \frac{S_n(\alpha)}{n} + \frac{\lambda_n p_n(\alpha)}{n}\frac{S_n(\alpha)}{n - p_n(\alpha)} = \frac{S_n(\alpha)}{n}\left[1 + \frac{\lambda_n p_n(\alpha)}{n - p_n(\alpha)}\right]. \tag{2.79}$$

Therefore,

$$\log[\Gamma_{n,\lambda_n}^{G_2}(\alpha)] = \log\frac{S_n(\alpha)}{n} + \log\left[1 + \frac{\lambda_n p_n(\alpha)}{n - p_n(\alpha)}\right].$$

It is our assumption that $p_n = O(n^\theta)$ with $\theta < 1/4$ and $\lambda_n p_n/n \to 0$ as $n \to \infty$. Using Taylor expansion,

$$\begin{aligned}
\log\left[1+\frac{\lambda_n p_n(\alpha)}{n-p_n(\alpha)}\right] &= \lambda_n \frac{p_n(\alpha)}{n-p_n(\alpha)} + O\left[\left(\frac{\lambda_n p_n(\alpha)}{n-p_n(\alpha)}\right)^2\right] \\
&= \lambda_n\left[\left(1-\frac{p_n(\alpha)}{n}\right)^{-1}-1\right] + o\left(\frac{\lambda_n p_n}{n}\right) \\
&= \lambda_n\left[\frac{p_n(\alpha)}{n} + o\left(\frac{p_n}{n}\right)\right] + o\left(\frac{\lambda_n p_n}{n}\right) \\
&= \frac{\lambda_n p_n(\alpha)}{n} + o\left(\frac{\lambda_n p_n}{n}\right), \qquad (2.80)
\end{aligned}$$

and therefore

$$\log[\Gamma^{G_2}_{n,\lambda_n}(\alpha)] = \log\frac{S_n(\alpha)}{n} + \frac{\lambda_n p_n(\alpha)}{n} + o\left(\frac{\lambda_n p_n}{n}\right).$$

Therefore

$$\log[\Gamma^{G_2}_{n,\lambda_n}(\alpha)] - AIC = o(p_n/n)$$

and

$$\log[\Gamma^{G_2}_{n,\lambda_n}(\alpha)] - BIC = o(1).$$

When $n \to \infty$, the criterion they minimize are close enough, so the methods are equivalent.

Let

$$\tilde{r} = \frac{\sigma_u^2+\sigma_e^2}{b\sigma_u^2+\sigma_e^2} < 1.$$

Under Assumption 2.3 , if we use s^2 in (2.50), then according to Theorem 2.11, the number of extra variables chosen by minimizing GIC_{λ_n} is

$$E[q_o] = q_n P_a^{G_1} = 2q_n[1-\Phi(\sqrt{\lambda_n \tilde{r}})] + O(n^{\theta-\frac{\delta}{2}}).$$

Approximately,

- C_p and FPE_λ with $\lambda > 2$ a constant: When λ_n in (2.50) is a constant, so is $P_a^{G_1}$ and therefore $E[q_o] = O(n^\theta)$, the expected number of extra variables is going to infinity at the same rate as p_n.

- Rao and Wu's GIC: Rao and Wu's GIC is the special case when we use s^2 as $\hat{\sigma}^2(\alpha)$ and $\lambda_n = C_n$, where C_n goes to infinity slower than n but faster than $\log\log n$. The sufficient condition for $E[q_o] \to 0$ when $n \to \infty$, i.e. for the GIC to choose a model that contains only the p^* true fixed effects, is that C_n grows at least as fast as $\log n$. To see this, note that for $C_n = c\log n$,

$$q_n(1-\Phi(\sqrt{\tilde{r}C_n})) = q_n\frac{e^{-\frac{\tilde{r}C_n}{2}}}{\sqrt{2\pi\tilde{r}C_n}}(1+O(C_n^{-2})) = \frac{an^{\theta-\frac{c\tilde{r}}{2}}}{\sqrt{2\pi\tilde{r}c\log n}}(1+O([\log(n)]^{-2})).$$

and since $s^2 = (\sigma_u^2+\sigma_e^2) + O_p(n^{-1/2})$,

$$q_n\left|\Phi(\sqrt{\tilde{r}C_n}) - \Phi\left(\sqrt{\frac{\tilde{r}s^2}{b\sigma_u^2+\sigma_e^2}}\right)\right| \le O(n^\theta)\sqrt{\log n}O_p(n^{-1/4}) \to 0.$$

Therefore

$$E[q_o] = 2q_n[1-\Phi(\sqrt{C_n\tilde{r}})(1+o(1))$$

and

  - The expected number of extra variables $E[q_o] \to 0$ when $\theta - \frac{\tilde{r}c}{2} \le 0$.
  - The expected number of extra variables $E[q_o] \to \infty$ when $\theta - \frac{\tilde{r}c}{2} > 0$.

If C_n grows to infinity slower than $\log n$, then $E[p_o] \to \infty$, and the faster C_n grows, the slower $E[p_o]$ grows.

Let

$$\phi_K \equiv 2\left[1-\Phi\left(\sqrt{\frac{K(\sigma_u^2+\sigma_e^2)}{b\sigma_u^2+\sigma_e^2}}\right)\right]. \tag{2.81}$$

Under Assumption 2.4, by Theorem 2.12 the expected number of extra variables in a sequential model selection

- C_p and FPE_λ with $\lambda > 2$ a constant:

$$E[p_o^{G_1}] = \phi_K(1 - \phi_K^{q_n})/(1 - \phi_K) + O(n^{-\delta}) \to \phi_K/(1 - \phi_K)$$

  where ϕ_K is defined in (2.81).

- Rao and Wu's GIC: When $\lambda_n \to \infty$, $P_{G_1} \to 0$ and

$$E[p_o^{G_1}] = P_{G_1}(1 - [P_{G_1}]^{q_n})/(1 - P_{G_1}) + O(n^{-\delta}) \to 0$$

  when $n \to \infty$.

Remark: From the above, we can see that λ_n has to grow to infinity fast enough to make GIC_{λ_n} choose a parsimonious model in balanced-data, orthogonal design; but in sequential selection, as long as $\lambda_n \to \infty$, asymptotically, the expected number of extra variables will be 0. □

2.5 Conclusions

We discussed the effect of omitting the random intercept in linear models. In this special simple model class, the MLE $\hat{\boldsymbol{\beta}}_n$ under the working model is still consistent in the sense that the Euclidean norm of the difference $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|$ converges to zero in probability. Therefore omitting the random effect does not give us wrong estimates for the parameter estimates, and should not give us spurious variables in hypothesis testing or model selection. On the other hand, because the model fails

to correctly specify the variance structure of the data, it is not surprising to find the estimated variance structure of the $\hat{\boldsymbol{\beta}}_n$ is not correct. The main reason that the inferences are unreliable when it comes to hypothesis testing and automatic model selection is because of the nonconsistent variance estimator. But this can be fixed by adopting a robustified variance estimator (2.32). The problem of increasing dimension is another reason to see spurious variables. When the dimension of the parameter space is fixed, a nonconsistent estimator for the variance structure will produce wrong inferences about the coefficient estimates in hypothesis testing and automatic model selection, but will not give us a number of spurious variables that goes to infinity. When the dimension is increasing with n, with the same probability of making a mistake, we will have infinitely many number of spurious variables due to failure to estimate the variance structure consistently. We therefore recommend that especially when the data have a clustered structure, to use the robustified variance estimator because even when the model is not correctly specified, this can still lead to correct statistical inferences, and avoid fitting a mixed effect model, which is more computationally burdensome than a fixed effect model.

Chapter 3

Generalized Linear Models

Generalized linear models(McCullagh and Nelder [17]) are used for regression analysis in a number of cases, including categorical data, where the classical assumption on normality of the data are violated. The statistical analysis of such models is based on the asymptotic large-sample properties of the maximum likelihood estimator. In this chapter we present the conditions that the MLE converges to a well-defined limit and is asymptotically normal under our design matrix assumptions, and will demonstrate the results with two special cases: the Logistic model and the Poisson model.

3.1 Notations

In this chapter we assume that the working model is a fixed-effect generalized linear model (GLM), while the true model contains a random effect for each cluster (GLMM).

The True Model: The response vector $\mathbf{y}$ is assumed to consist of independent elements conditional on the random-effect vector $\mathbf{u}$, each with a distribution with density from the exponential family:

$$y_{ij}|\mathbf{u} \overset{\text{indep.}}{\sim} f_{Y_{ij}|\mathbf{u}}(y_{ij}|\mathbf{u}), \quad i = 1, \cdots, m, \quad j = 1, \cdots, n_i$$

$$f_{Y_{ij}|\mathbf{u}}(y_{ij}|\mathbf{u}) = \exp\{[y_{ij}\gamma_{ij}^* - b^*(\gamma_{ij}^*)]/\tau^{*2} - c^*(y_{ij}, \tau^*)\}. \tag{3.1}$$

We model the transformation of the conditional mean of y_{ij} as a linear model in both the fixed and random effects:

$$E[y_{ij}|\mathbf{u}] = \mu^*_{ij} \tag{3.2}$$

$$g^*(\mu^*_{ij}) = \mathbf{x}^*_{ij}\boldsymbol{\beta}^* + u_i$$

Here the link function $g^*(\cdot)$ is assumed known, $\mathbf{x}^*_{ij}$ is the $(\sum_{k=1}^{i-1} n_k + j)^{th}$ row of the model matrix for the fixed effects corresponding to the response y_{ij}, and $\boldsymbol{\beta}^*$ is the $p^* \times 1$ fixed effects parameter vector. The parameter γ^*_{ij} is in an open region in $\mathbf{R}$ and is related to $\mathbf{x}^*_{ij}\boldsymbol{\beta}^*$ through

$$\frac{\partial b(\gamma^*_{ij})}{\partial \gamma^*_{ij}} = \mu^*_{ij} = g^{-1}(\mathbf{x}^*_{ij}\boldsymbol{\beta}^* + u_i).$$

To that specification we have added $\mathbf{u}$, the random effects vector. Finally, we assume that the random effect u_i's are iid from known density $f_U(u)$ for $i = 1, \cdots, m$.

The Working Model: The vector $\mathbf{y}$ is assumed to consist of independent measurements from a distribution with density from the exponential family:

$$y_{ij} \overset{\text{indep.}}{\sim} \quad f_{Y_{ij}}(y_{ij}), \quad i = 1, \cdots, m \quad j = 1, \cdots, n_i$$

$$f_{Y_{ij}}(y_{ij}) = \exp\{[y_{ij}\gamma_{ij} - b(\gamma_{ij})]/\tau^2 - c(y_{ij}, \tau)\}. \tag{3.3}$$

The link function $g(\cdot)$ relates the transformation of the mean, μ_{ij}, as a linear model in the predictors:

$$E[y_{ij}] = \mu_{ij} = \frac{\partial b(\gamma_{ij})}{\partial \gamma_{ij}}.$$

$$g(\mu_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}_n, \tag{3.4}$$

where $g(\cdot)$ is a known function, $\mathbf{x}_{ij}$ is the double-indexed row of the model matrix corresponding to the response y_{ij}, and $\boldsymbol{\beta}_n$ is the $p_n \times 1$ parameter vector in the linear predictor. Again the parameter γ_{ij} is related to $\mathbf{x}_{ij}\boldsymbol{\beta}$ through

$$\frac{\partial b(\gamma_{ij})}{\partial \gamma_{ij}} = \mu_{ij} = g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_n).$$

In this chapter we assume that the covariates are designed within-cluster covariates, or $\tilde{\mathbf{X}}_n = \mathbf{X}_n$ with all rows stochastically independent. We will therefore use the notation $\mathbf{X}_n$ only.

3.2 The General Model

Conditions to assure consistency and asymptotic normality of the MLE for exponential families have previously been discussed in the literature under regularity conditions. See Berk [2], Fahrmeir and Kaufmann [10] for detailed discussions. There are also discussions of these conditions with parameters of increasing dimensions. Portnoy [20] discussed consistency and asymptotic normality for the MLE of the population mean of the exponential family when the number of parameters tends to infinity, and He and Shao [12]considered M-estimators of general parametric models with expanding dimensions. Both authors gave the rate at which p_n is allowed to grow for the asymptotic distributional approximations of the estimators (MLE for Portnoy) to be still valid, but Portnoy did not consider the case where the model is misspecified while Shao and He's conditions are not easy to check even

with the logistic model in our case; Strawderman and Tsiatis [27] applied the Inverse Function Theorem to get consistency results when the parameter space is expanding and the model could be misspecified, but their conditions are too restrictive for us to apply. White [32] discussed the asymptotic properties of the MLE under model misspecification, but only considered the case where the parameter space is fixed, and the data are iid. Our special situation cannot be an application of any of these past discussions. But in this chapter, we can use methods that are similar to those used in White [32] and Portnoy [20] to get our own conditions for consistency and asymptotic normality of the MLE under the working model.

3.2.1 The Likelihood Equations

We will first look at the likelihood equations, which are discussed in various textbooks. We adopt the notations in McCulloch and Searle [18]. The log likelihood for the working model (3.3) is given by

$$l_n(\boldsymbol{\beta}_n) = \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} [y_{ij}\gamma_{ij} - b(\gamma_{ij})]/\tau^2 - \sum_{i=1}^{m} \sum_{j=1}^{n_i} c(y_{ij}, \tau) \right). \tag{3.5}$$

Define

$$g_\mu(\mu_{ij}) = \frac{\partial g(\mu_{ij})}{\partial \mu_{ij}},$$

$$v(\mu_{ij}) = \frac{\partial^2 b(\gamma_{ij})}{\partial \gamma_{ij}^2},$$

$$w_{ij} = [v(\mu_{ij}) g_\mu^2(\mu_{ij})]^{-1},$$

and use two very useful identities in generalized linear models:

$$\frac{\partial \gamma_{ij}}{\partial \mu_{ij}} = \left(\frac{\partial \mu_{ij}}{\partial \gamma_{ij}} \right)^{-1} = \left(\frac{\partial^2 b(\gamma_{ij})}{\partial \gamma_{ij}^2} \right)^{-1} = \frac{1}{v(\mu_{ij})}; \tag{3.6}$$

and

$$\begin{aligned}\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_n} &= \frac{\partial \mu_{ij}}{\partial g(\mu_{ij})}\frac{\partial g(\mu_{ij})}{\partial \boldsymbol{\beta}_n} = \left(\frac{\partial g(\mu_{ij})}{\partial \mu_{ij}}\right)^{-1}\frac{\partial \mathbf{x}_{ij}\boldsymbol{\beta}_n}{\partial \boldsymbol{\beta}_n} \\ &= \left(\frac{\partial g(\mu_{ij})}{\partial \mu_{ij}}\right)^{-1}\mathbf{x}'_{ij}. \end{aligned} \tag{3.7}$$

Then we have

$$\begin{aligned}\frac{\partial l_n(\boldsymbol{\beta}_n)}{\partial \boldsymbol{\beta}_n} &= \frac{1}{\tau^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\frac{\partial \gamma_{ij}}{\partial \boldsymbol{\beta}_n} - \frac{\partial b(\gamma_{ij})}{\partial \gamma_{ij}}\frac{\partial \gamma_{ij}}{\partial \boldsymbol{\beta}_n}\right] \\ &\overset{(3.2)}{=} \frac{1}{\tau^2}\sum_i\sum_j (y_{ij}-\mu_{ij})\frac{\partial \gamma_{ij}}{\partial \boldsymbol{\beta}_n} \\ &= \frac{1}{\tau^2}\sum_i\sum_j (y_{ij}-\mu_{ij})\frac{\partial \gamma_{ij}}{\partial \mu_{ij}}\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_n} \\ &\overset{(3.6)(3.7)}{=} \sum_i\sum_j \frac{(y_{ij}-\mu_{ij})}{v(\mu_{ij})g_\mu(\mu_{ij})}\mathbf{x}'_{ij} \\ &= \frac{1}{\tau^2}\sum_i\sum_j (y_{ij}-\mu_{ij})w_{ij}g_\mu(\mu_{ij})\mathbf{x}'_{ij}. \end{aligned} \tag{3.8}$$

We can write this in matrix notation as

$$\frac{\partial l_n(\boldsymbol{\beta}_n)}{\partial \boldsymbol{\beta}_n} = \frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}), \tag{3.9}$$

with

$$\mathbf{W}_{n\times n} = \text{diag}(w_{ij}) \text{ and } \boldsymbol{\Delta}_{n\times n} = \text{diag}(g_\mu(\mu_{ij})), \text{ for } j \in \{1,\cdots,n_i\},\, i \in \{1,\cdots,m\}. \tag{3.10}$$

The ML equations are thus given by

$$\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}\mathbf{y} = \mathbf{X}'\mathbf{W}\boldsymbol{\Delta}\boldsymbol{\mu}, \tag{3.11}$$

where $\mathbf{W}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\mu}$ involve the unknown $\boldsymbol{\beta}_n$. Note that the MLE of $\boldsymbol{\beta}_n$ remains the same in the presence of the nuisance parameter τ. Typically these are not linear functions of $\boldsymbol{\beta}_n$ and so cannot be solved analytically.

3.2.2 Consistency of MLE under Nonstandard Conditions

The usual regularity conditions on the existence of a unique and consistent solution to the likelihood equations include that the underlying probability distribution of the data is a member of the parametric family considered, and that the dimension of the parameter is fixed. These conditions are not satisfied in our case. By studying the working likelihood at the cluster level, we can use large-sample asymptotics to draw conclusions analogous to that of White [32] when the parameter space has a fixed dimension.

Suppose that the parameter space $\mathcal{B}_n$ for $\boldsymbol{\beta}_n$ is an open region in $\mathbf{R}^{p_n}$. Let $G_n(\boldsymbol{\beta}_n) \equiv m^{-1} l_n(\boldsymbol{\beta}_n)$ be the normalized log likelihood, and $\boldsymbol{\beta}_n^\star$ be the solution to $E\left[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)\right] = \mathbf{0}$; then we have

Theorem 3.1 *Suppose that there exists $\epsilon > 0$ independent of n such that*

1. *$G_n(\boldsymbol{\beta}_n)$ is a concave function of $\boldsymbol{\beta}_n$,*
2. *$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\| = o_p(1)$;*
3. *$\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) \leq -C\mathbf{I}_{p_n}$ for all $\boldsymbol{\beta}_n \in \mathcal{B}_\epsilon(\boldsymbol{\beta}_n^\star)$, with probability approaching 1 as $n \to \infty$, for a constant C that could depend on ϵ but not on n.*

Then as $n \to \infty$, with probability approaching to 1 there exists a unique solution $\hat{\boldsymbol{\beta}}_n \in \mathcal{B}_\epsilon(\boldsymbol{\beta}_n^\star)$ to the equation $\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) = \mathbf{0}$, and $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star\| = O_p(\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\|)$.

Proof: Any continuous function has a local minimum or maximum in a compact set, and for concave functions, a local maximum is also the global maximum. If we can prove that outside the $\epsilon-$neighborhood of $\boldsymbol{\beta}_n^\star$ there cannot be a maximizer

of $G_n(\boldsymbol{\beta}_n)$, then within the compact $\epsilon-$neighborhood of $\boldsymbol{\beta}_n^\star$ the concave function $G_n(\boldsymbol{\beta}_n)$ must have a unique maximum and the theorem is proved.

For $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\| \leq \epsilon$, the following Taylor expansion (in Mean Value Theorem form) holds for some $\tilde{\boldsymbol{\beta}}_n$ between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_n^\star$:

$$G_n(\boldsymbol{\beta}_n) - G_n(\boldsymbol{\beta}_n^\star) = \left(\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right)'(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star) + \frac{1}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star)'\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star).$$

By Condition 2,

$$(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star)'\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star) \leq -C(\boldsymbol{\beta} - \boldsymbol{\beta}_n^\star)'(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star) = -C\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\|^2.$$

Therefore, there exists a sequence $\alpha_n = o_p(1)$ such that

$$G_n(\boldsymbol{\beta}_n) - G_n(\boldsymbol{\beta}_n^\star) \leq \alpha_n\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\| - \frac{C}{2}\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\|^2 < 0 \tag{3.12}$$

for $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\| \geq 2\alpha_n/C$. For n large enough $2\alpha_n/C < \epsilon$, and $G_n(\boldsymbol{\beta}_n) - G_n(\boldsymbol{\beta}_n^\star) < 0$ as long as $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\| > \epsilon$. This means that the maximum cannot be outside of the $\epsilon-$neighborhood of $\boldsymbol{\beta}_n^\star$. Within the compact set $\mathcal{B}_\epsilon(\boldsymbol{\beta}_n^\star)$, the concave function $G_n(\boldsymbol{\beta}_n)$ has a maximizer $\hat{\boldsymbol{\beta}}_n$ which solves $\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) = \mathbf{0}$, and $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star\| = O_p\left(\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\|\right) \xrightarrow{p} 0$. □

Since

$$E[y_{ij}|\mathbf{x}_{ij}] = E_u[g^{*-1}(\mathbf{x}_{ij}^*\boldsymbol{\beta}^* + u)] = \int g^{*-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u) f_U(u)du,$$

it follows that

$$\begin{aligned} E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] &= \frac{1}{m\tau^2} E_{(\mathbf{x}_i, n_i)} \sum_i \sum_j \left[\left(E[y_{ij}|\mathbf{x}_{ij}] - \mu_{ij} w_{ij} g_\mu(\mu_{ij})\right)\mathbf{x}_{ij}'\right] \\ &= \frac{1}{m\tau^2} E_{(\mathbf{x}_i, n_i)} \sum_i \sum_j \Big[\Big(E_u[g^{*-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u)] \\ &\quad - g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_n) w(\mathbf{x}_{ij}\boldsymbol{\beta}_n) g_\mu(\mathbf{x}_{ij}\boldsymbol{\beta}_n)\Big)\mathbf{x}_{ij}'\Big] \end{aligned} \tag{3.13}$$

where the subscript in the expectation denotes the variables over which the expectation is taken. We write w and g_μ as functions of $\mathbf{x}_{ij}\boldsymbol{\beta}_n$ to emphasize the fact that both of them depend on $\mathbf{x}_{ij}\boldsymbol{\beta}_n$ and cannot be written out of the expectation. The expectation is taken under the joint density of $\mathbf{x}_{ij}$ and n_i. For simplicity of the notations we further assume that

Assumption 3.1 $\mathbf{x}_{ij}$ *are independent of* n_i *for all* i *and* j.

Under this Assumption, (3.13) becomes

$$E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] = \frac{1}{\tau^2} E_{\mathbf{x}} \left[\left(E_u[g^{*-1}(\mathbf{x}\boldsymbol{\beta}_0 + u) - g^{-1}(\mathbf{x}\boldsymbol{\beta}_n) w(\mathbf{x}\boldsymbol{\beta}_n) g_\mu(\mathbf{x}\boldsymbol{\beta}_n) \right) \mathbf{x}' \right].$$

The solution to (3.13), $\boldsymbol{\beta}_n^\star$, therefore satisfies

$$E_{\mathbf{x}} \left[\left(\int g^{*-1}(\mathbf{x}\boldsymbol{\beta}_0 + u) f_U(u) du - g^{-1}(\mathbf{x}\boldsymbol{\beta}_n^\star) w(\mathbf{x}\boldsymbol{\beta}_n^\star) g_\mu(\mathbf{x}\boldsymbol{\beta}_n^\star) \right) \mathbf{x}' \right] = \mathbf{0}.$$

Remark 1: This result is analogous to that of White [32] because $\boldsymbol{\beta}_n^\star$ in Theorem 3.1 is actually the parameter in $\mathcal{B}_n$ that minimizes the Kullback-Leibler [14] information. Under the model misspecification and assumptions we impose on the regressors $\mathbf{x}_{ij}$, we have proved that the MLE of the working model (Quasi-MLE in White's discussion [32]) converges in probability to a well defined limit. □

Remark 2: We impose Assumption 3.1 only to reduce the formula to a more interpretable form; to have the conditions on $G_n(\boldsymbol{\beta}_n^\star)$ and its derivatives satisfied, we only need moments of $\mathbf{x}'_{ij}$ to be well defined, and as n_i's are bounded almost surely, conditions on expectation taken with respect to the joint density is virtually the same as those on expectation taken with respect to the density of $\mathbf{x}_{ij}$ alone. Therefore without loss of generality we will continue to impose Assumption 3.1 throughout this chapter.

3.2.3 Asymptotic Normality

When the parameter space has fixed dimension, under regularity conditions (Berk [2]) the MLE is asymptotically normal with $p \times 1$ mean and $p \times p$ variance matrix. Even when the model is not correctly specified, the asymptotic normality can still be established when the number of parameters is fixed (White [32]). Since each $\hat{\boldsymbol{\beta}}_n$ is a $p_n \times 1$ vector, with p_n increasing with n, we need another form of asymptotic normality. Definition (1.12) in Chapter 1 is a strong version of normality in the Central Limit Theorem that was used both in Portnoy [20] and Shao and He [12], since apparently it implies the element-wise normality of $\hat{\boldsymbol{\beta}}_n$.

First notice that the gradient $\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G(\boldsymbol{\beta}_n^\star)$ can be written as:

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) = \frac{1}{m\tau^2} \sum_i \sum_j [y_{ij} - \mu_{ij} w_{ij} g_\mu(\mu_{ij})] \mathbf{x}'_{ij} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\zeta}_i,$$

where under the true model the $p_n \times 1$ random vectors

$$\boldsymbol{\zeta}_i \equiv \sum_{j=1}^{n_i} (y_{ij} - \mu_{ij} w_{ij} g_\mu(\mu_{ij})) \mathbf{x}'_{ij} \tag{3.14}$$

are iid with $E\boldsymbol{\zeta}_i = \mathbf{0}$ and

$$E\left[\boldsymbol{\zeta}_i^{\otimes 2}\right] = mE\left[\left(\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right)^{\otimes 2}\right] \tag{3.15}$$

Then we can take the same approach as White [32] to establish the asymptotic normality of $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star)$:

Theorem 3.2 *Under Assumptions 1.2-1.4, if there exists a $\epsilon > 0$ not shrinking with n such that for all $\boldsymbol{\beta}_n \in \mathcal{B}_\epsilon(\boldsymbol{\beta}_n^\star)$*

1. *$G_n(\boldsymbol{\beta}_n)$ is a concave function for $\boldsymbol{\beta}_n$;*

2. $\left\| \nabla_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \right\| = O_p(p_n/\sqrt{n})$;

3. *For a constant C that could depend on ϵ but not on n,*

$$\nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) \leq -C\mathbf{I}_{p_n}$$

and there exists a $\delta_1 > 0$ such that

$$\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^\star\| \leq \epsilon} \left\| \left(\nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) \right)^{-1} - \left(\nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \right)^{-1} \right\| = O_p(p_n^{-1-\delta_1});$$

4. *There exists a $\delta_2 > 0$ such that*

$$\left\| \left(E\left[\nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \right]^{-1} \right) - \left(\nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \right)^{-1} \right\| = O_p(p_n^{-1-\delta_2}) \tag{3.16}$$

5. *For any sequence of unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$ the sequence $\mathbf{v}_n' \boldsymbol{\zeta}_i$, $1 \leq i \leq m$ satisfy the Lyapunov condition for central limit theorem.*

Then for any sequence of unit vectors $\mathbf{v}_n \in \mathbf{R}^{p_n}$ with $\sigma^2_{\mathbf{v}_n} = \mathbf{v'}_n \mathbf{A}_n^{-1} \mathbf{B}_n \mathbf{A}_n^{-1} \mathbf{v}_n$,

$$\sqrt{m}\sigma^{-1}_{\mathbf{v}_n} \mathbf{v}'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star) \to \mathcal{N}(0,1).$$

Here

$$\mathbf{A}_n = -E\left[\nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \right]$$

and

$$\mathbf{B}_n = E\left[\boldsymbol{\zeta}_i^{\otimes 2} \right]$$

where the expectations are taken under the true model.

Proof: Since $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star\| = O_p(p_n/\sqrt{n})$, the Taylor expansion of the function $\nabla_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)$ at $\hat{\boldsymbol{\beta}}_n$ yields

$$\mathbf{0} = \nabla_{\boldsymbol{\beta}_n} G_n(\hat{\boldsymbol{\beta}}_n) = \nabla_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) + \nabla^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star)$$

for some $\tilde{\boldsymbol{\beta}}_n$ between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_n^\star$ by Mean Value Theorem. By Condition 3,

$$-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n) \geq C\mathbf{I}_{p_n}$$

uniformly in the $\epsilon-$neighborhood of $\boldsymbol{\beta}_n^\star$, so $-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)$ is invertible in the $\epsilon-$neighborhood of $\boldsymbol{\beta}_n^\star$, and

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star &= \left(-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)\right)^{-1} \boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \\
&= \left(E\left[-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right]\right)^{-1} \boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \\
&\quad + \left\{\left(-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)\right)^{-1} - \left(E\left[-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right]\right)^{-1}\right\} \boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star).
\end{aligned}$$

Uniformly for any unit vector $\mathbf{v}_n \in \mathbf{R}^{p_n}$,

$$\begin{aligned}
&\mathbf{v}'_n \left\{\left(-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)\right)^{-1} - \left(E\left[-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right]\right)^{-1}\right\} \boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) \\
\leq\ & \left\|\left(-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)\right)^{-1} - \left(E\left[-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right]\right)^{-1}\right\| \left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\| \\
\leq\ & \left\|\left(E\left[\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right]\right)^{-1} - \left(\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right)^{-1}\right\| \left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\| \\
&+ \left\|\left(\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\tilde{\boldsymbol{\beta}}_n)\right)^{-1} - \left(\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right)^{-1}\right\| \left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\| \\
=\ & O_p(n^{-1/2}p_n^{-\delta_1}) + O_p(n^{-1/2}p_n^{-\delta_2}). \qquad (3.17)
\end{aligned}$$

Therefore, uniformly for any unit vector $\mathbf{v}_n \in \mathbf{R}^{p_n}$,

$$\sqrt{m}\mathbf{v}'_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star) = \frac{1}{\sqrt{m}}\mathbf{v}'_n\mathbf{A}_n^{-1}\sum_{i=1}^m \boldsymbol{\zeta}_i + O_p(p_n^{-\delta_1}) + O_p(p_n^{-\delta_2}).$$

By Condition 5

$$\frac{1}{\sqrt{m}}\tilde{\sigma}^{-1}_{\mathbf{v}_n}\sum_{i=1}^m \mathbf{v}'_n\boldsymbol{\zeta}_i \to \mathcal{N}(0,1)$$

where

$$\tilde{\sigma}^2_{\mathbf{v}_n} = \mathbf{v}'_n E\left[\boldsymbol{\zeta}_i^{\otimes 2}\right]\mathbf{v}_n.$$

Finally, similar to the proof of Theorem 2.3 in Chapter 2, $\|\mathbf{A}_n^{-1}\|$ is bounded and $\mathbf{v}_n'\mathbf{A}_n^{-1}\boldsymbol{\zeta}_i$ satisfies the Lyapunov condition. So

$$\sqrt{m}\sigma_{\mathbf{v}_n}^{-1}\mathbf{v}'_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star) \to \mathcal{N}(0,1)$$

where $\sigma_{\mathbf{v}_n}^2 = \mathbf{v}'_n\mathbf{A}_n^{-1}E\left[\boldsymbol{\zeta}_i^{\otimes 2}\right]\mathbf{A}_n^{-1}\mathbf{v}_n$. □

Remark 1: Theorem 3.2 gives a result analogous to that of ordinary linear regression models, namely, a robust estimator for the variance of the MLE. Because the random effect was omitted in the working model, part of the variability of the data is not explained by the working model, and the usual working-model variance estimator for $\hat{\boldsymbol{\beta}}_n$ will be biased. Aside from the fact that $\hat{\boldsymbol{\beta}}_n$ is not necessarily consistent ($\|\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0\| \not\to 0$), the bias in estimating the variance of $\hat{\boldsymbol{\beta}}_n$ could lead to unreliable statistical inferences as we have seen in Section 2.3 and 2.4. We will see in the computational part of the discussion (Section 3.5) that the "sandwich" variance estimator estimates the variance of $\hat{\boldsymbol{\beta}}_n$ very well, even when the model is misspecified. □

Remark 2: The assumptions will be verified in particular models in later sections of this Chapter. □

3.3 Logistic Regression: A Special Case

In this section, logistic model as a special case of generalized linear models is studied. There are many desirable features about the logistic model: the data are bounded therefore have infinitely many finite moments; the natural link function makes the log likelihood a linear function of y_{ij}, so that the Hessian of $l_n(\boldsymbol{\beta}_n)$ does not

involve y_{ij}; and the Hessian matrix of $l_n(\boldsymbol{\beta}_n)$ is negative definite so $l_n(\boldsymbol{\beta}_n)$ is concave in $\boldsymbol{\beta}_n$. Therefore $\boldsymbol{\beta}_n^\star$ is unique if it exists. For the moment we only consider the case where the link function is correctly specified, namely, the case where $g(\cdot) = g^*(\cdot)$, $b(\cdot) = b^*(\cdot)$ and $c(\cdot) = c^*(\cdot)$ in (3.1) and (3.3). In this section, we are going to discuss in detail what to expect of $\boldsymbol{\beta}_n^\star$ when σ_u^2, the variance of u_i, is in different ranges.

3.3.1 Notations and Assumptions

The Logistic-Normal model makes the following assumption about the density of the random effect:

Assumption 3.2 *The random effects, u_i, are i.i.d. normal variates with mean* 0 *and variance σ_u^2.*

The log likelihood of the working model with sample size n is therefore

$$\begin{aligned} l_n(\boldsymbol{\beta}_n) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left(\mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}\right) \\ &= \frac{1}{n}\sum_i\sum_j \left[y_{ij}\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) + \log(1-\mu_{ij})\right], \end{aligned} \tag{3.18}$$

where

$$\mu_{ij} = Ey_{ij} = P[y_{ij} = 1]$$

and both the expectation and the probability are taken under the working model.

The GLM notations for this case are

$$\tau^2 = 1,$$

$$\gamma_{ij} = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right),$$

$$b(\gamma_{ij}) = -\log(1-\mu_{ij}) = \log(1+e^{\gamma_{ij}}),$$

and

$$c(y_{ij}, \tau) = 0.$$

Using the canonical link

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),$$

(3.18) becomes

$$l_n(\boldsymbol{\beta}_n) = \sum_i \sum_j \left[y_{ij}\mathbf{x}_{ij}\boldsymbol{\beta}_n - \log(1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}) \right]. \tag{3.19}$$

3.3.2 Asymptotic Limit of $\hat{\boldsymbol{\beta}}_n$

First of all, with the working model in mind, we can see that

$$G_n(\boldsymbol{\beta}_n) = \frac{1}{m} l_n(\boldsymbol{\beta}_n) = \frac{1}{m} \sum_i \sum_j \left[y_{ij}\mathbf{x}_{ij}\boldsymbol{\beta}_n - \log(1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}) \right],$$

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) = \frac{1}{m} \boldsymbol{\nabla}_{\boldsymbol{\beta}_n} l_n(\boldsymbol{\beta}_n) = \frac{1}{m} \sum_i \sum_j \left[y_{ij} - \frac{e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}}{1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}} \right] \mathbf{x}'_{ij}, \tag{3.20}$$

and

$$\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) = \frac{1}{m} \boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} l_n(\boldsymbol{\beta}_n) = -\frac{1}{m} \sum_i \sum_j \frac{e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}}{(1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n})^2} \mathbf{x}'_{ij}\mathbf{x}_{ij}. \tag{3.21}$$

Therefore $\boldsymbol{\beta}^\star_n$ under Assumptions 3.1 and 3.2 satisfies

$$E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}^\star_n)] = E_{\mathbf{x}} \left[\left(\int \frac{e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}} \phi(z) dz - \frac{e^{\mathbf{x}\boldsymbol{\beta}^\star_n}}{1+e^{\mathbf{x}\boldsymbol{\beta}^\star_n}} \right) \mathbf{x}' \right] = \mathbf{0}, \tag{3.22}$$

where $\phi(\cdot)$ is the pdf of a standard normal distribution.

Theorem 3.3 *Under Assumptions 1.2-1.4, as $n \to \infty$, a unique solution $\hat{\boldsymbol{\beta}}_n$ of the equation $\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}$ exists in a fixed neighborhood $\mathcal{B}_\epsilon(\boldsymbol{\beta}_n^\star)$ about $\boldsymbol{\beta}_n^\star$ with probability going to one, and $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star\| = O_p(\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\|)$.*

Proof: By (3.14) and (3.20),

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\zeta}_i \equiv \frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{n_i}\left(y_{ij} - \frac{e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star}}{1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star}}\right)\mathbf{x}_{ij}'\right].$$

where the iid $p_n \times 1$ random vectors $\boldsymbol{\zeta}_i$ satisfies $E[\boldsymbol{\zeta}_i] = \mathbf{0}$. Let $\boldsymbol{\zeta}_{ik}$ be the k^{th} entry of $\boldsymbol{\zeta}_i$, then by Assumption 1.3 and the fact that $\left[y_{ij} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}(1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n})^{-1}\right]$ is bounded for all i, j and $\boldsymbol{\beta}_n \in \mathcal{B}_n$, we have $E|\boldsymbol{\zeta}_{ik}|^{4r} < \infty$ uniformly in $1 \le k \le p_n$. Therefore for any constant $K > 0$,

$$\begin{aligned}
& \sqrt{p_n} P\left[\max_{1\le k\le p_n}\left|\frac{1}{n}\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right| \ge \frac{Kp_n}{\sqrt{n}}\right] \\
\le\ & p_n^{3/2} P\left[\left|\frac{1}{n}\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right| \ge \frac{Kp_n}{\sqrt{n}}\right] \\
\le\ & M_r p_n^{3/2} n^{2r} p_n^{-4r} n^{-2r} K^{-4r} \qquad (3.23) \\
=\ & O(p_n^{3/2-4r}) \to 0
\end{aligned}$$

for $r > 1$, where (3.23) follows by Proposition B.1 in the Appendix. Therefore

$$\begin{aligned}
\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{m}\boldsymbol{\zeta}_i\right\| \\
&\le \sqrt{p_n}\max_{1\le k\le p_n}\left|\frac{1}{n}\sum_{i=1}^{m}\boldsymbol{\zeta}_{ik}\right| \\
&\le O_p(p_n/\sqrt{n}). \qquad (3.24)
\end{aligned}$$

For the logistic model, $G_n(\boldsymbol{\beta}_n)$ is always concave, and

$$-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star) = \frac{1}{m}\sum_i\sum_j \frac{e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star}}{(1+e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star})^2}\mathbf{x}_{ij}'\mathbf{x}_{ij} \le \frac{1}{4m}(\mathbf{X}_n'\mathbf{X}_n) \le \frac{1}{4}(M^* + a_n)\mathbf{I}_{p_n}$$

by Theorem 1.1. Therefore the conditions in Theorem 3.1 are met and $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star\right\|$ converges to zero in probability. □

Generally, though we can prove the existence of $\boldsymbol{\beta}_n^\star$ and the consistency of $\hat{\boldsymbol{\beta}}_n$, neither the likelihood equations or $E[S_n(\boldsymbol{\beta}_n)] = \mathbf{0}$ can be solved analytically. In the subsequent sections, we study the special cases where σ_u is in an extreme range and we can use Taylor expansion to get an approximation of $\boldsymbol{\beta}_n^\star$.

3.3.3 Asymptotic Normality of $\hat{\boldsymbol{\beta}}_n$

We will check the conditions in Theorem 3.2 in the logistic model to establish the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$. Define functions

$$p(x) = \frac{e^x}{1+e^x}$$

and

$$d(x) = \frac{e^x}{(1+e^x)^2}.$$

Theorem 3.4 *Under Assumptions 1.2-1.4, if* $r > 1/[2(1-3\theta)]$ *and*

$$\max_{\mathbf{v}_n \in \mathbf{R}^{p_n}: \|\mathbf{v}_n\|=1} E|\mathbf{x}_{ij}\mathbf{v}_n|^{2+\varepsilon} \leq C_\varepsilon < \infty$$

for some $\varepsilon > 0$, *then*

$$\sqrt{m}\sigma_{\mathbf{v}_n}^{-1}\mathbf{v}_n'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star) \to \mathcal{N}(0,1)$$

where $\sigma_{\mathbf{v}_n}^2 = \mathbf{v}_n'(\mathbf{A}_n)^{-1}\mathbf{B}_n(\mathbf{A}_n)^{-1}\mathbf{v}_n$,

$$\mathbf{A}_n = E\left[\mathbf{x}_{ij}' d(\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star)\mathbf{x}_{ij}\right]$$

and

$$\mathbf{B}_n = E\left(\sum_{j=1}^{n_i}(y_{ij} - p(\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star))\mathbf{x}_{ij}'\right)^{\otimes 2}.$$

Proof: Condition 1,2 and the first part of Condition 3 are easily checked by the proof of Theorem 3.3. What we need to prove for Condition 3 and 4 of Theorem 3.2 is that there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$\left\| \left(\mathbf{X}_n' \tilde{\mathbf{D}} \mathbf{X}_n \right)^{-1} - \left(\mathbf{X}_n' \mathbf{D}^\star \mathbf{X}_n \right)^{-1} \right\| = O_p(p_n^{-1-\delta_1}),$$

and

$$\left\| \left(\mathbf{X}_n' \mathbf{D}^\star \mathbf{X}_n \right)^{-1} - \left(E\left[\mathbf{X}_n' \mathbf{D}^\star \mathbf{X}_n \right] \right)^{-1} \right\| = O_p(p_n^{-1-\delta_2}).$$

where $\tilde{\mathbf{D}}$ and $\mathbf{D}^\star$ are $n \times n$ diagonal matrices defined by

$$\tilde{\mathbf{D}} \equiv m^{-1} \mathrm{diag}\left(d(\mathbf{x}_{11}\tilde{\boldsymbol{\beta}}_n), \cdots, d(\mathbf{x}_{mn_m}\tilde{\boldsymbol{\beta}}_n) \right)$$

and

$$\mathbf{D}^\star \equiv m^{-1} \mathrm{diag}\left(d(\mathbf{x}_{11}\boldsymbol{\beta}_n^\star), \cdots, d(\mathbf{x}_{mn_m}\boldsymbol{\beta}_n^\star) \right).$$

By Assumption 1.3, for $\boldsymbol{\beta}_n$ in a compact set in $\mathcal{B}_n$, $\mathbf{x}_{ij}\boldsymbol{\beta}_n$ is bounded almost surely, so there exist M_1 and M_2 such that

$$0 < M_1 < d(\mathbf{x}_{ij}\boldsymbol{\beta}_n) \leq M_2 < 1$$

for $\boldsymbol{\beta}_n$ in an ϵ-neighborhood of $\boldsymbol{\beta}_n^\star$. Therefore both $n^{-1}\mathbf{X}_n'\tilde{\mathbf{D}}\mathbf{X}_n$ and $n^{-1}\mathbf{X}_n'\mathbf{D}^\star\mathbf{X}_n$ are bounded below by $M_1(m^* - a_n)\mathbf{I}_{p_n}$ and above by $M_2(M^* + a_n)\mathbf{I}_{p_n}$ with probability approaching 1 as $n \to \infty$, and $n^{-1}E\left[\mathbf{X}_n'\mathbf{D}\mathbf{X}_n\right]$ is bounded below by $M_1 m^* \mathbf{I}_{p_n}$ and above by $M_2 M^* \mathbf{I}_{p_n}$. Therefore it suffices to prove that

$$\left\| \mathbf{X}_n' \mathbf{D}^\star \mathbf{X}_n - E\left[\mathbf{X}_n' \mathbf{D}^\star \mathbf{X}_n \right] \right\| = O_p(p_n^{-1-\delta_2}) \tag{3.25}$$

and

$$\left\| \mathbf{X}_n' \tilde{\mathbf{D}} \mathbf{X}_n - \mathbf{X}_n' \mathbf{D}^\star \mathbf{X}_n \right\| = O_p(p_n^{-1-\delta_1}) \tag{3.26}$$

for some positive numbers δ_1 and δ_2. By Theorem B.1, (3.25) is satisfied when $r > \theta/(1-4\theta)$ because $d(\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star)$ is bounded (and therefore has infinitely many finite moments). To see (3.26), note that $d(x)$ and its first derivative are both bounded and when $r > 1/(2(1-3\theta))$, for any $\delta_4 > 0$

$$\begin{aligned}
\left\|\mathbf{X}_n'\tilde{\mathbf{D}}\mathbf{X}_n - \mathbf{X}_n'\mathbf{D}^\star\mathbf{X}_n\right\| &\leq \frac{1}{\sqrt{n}}\left\|\tilde{\mathbf{D}} - \mathbf{D}^\star\right\|\sqrt{\frac{1}{n}\left\|\mathbf{X}_n'\mathbf{X}_n\right\|} \\
&\leq \sqrt{\frac{M^* + a_n}{n}}\max_{1\leq t\leq n}|d(\mathbf{x}_t\tilde{\boldsymbol{\beta}}_n) - d(\mathbf{x}_t\boldsymbol{\beta}_n^\star)| \\
&\leq C\sqrt{\frac{M^* + a_n}{n}}\max_t|\mathbf{x}_t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)| \\
&\leq C\sqrt{\frac{M^* + a_n}{n}}\sqrt{p_n}\max_{t,k}|\mathbf{x}_{tk}|\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \\
&\overset{Th.1.2}{=} O_p(n^{-1/2}p_n^{1/2}p_n n^{\frac{1}{4r}+\delta_4}p_n n^{-1/2}) \\
&= O_p(p_n^{1/2}n^{-1/2}) \qquad (3.27)
\end{aligned}$$

and therefore there exists some $\delta_2 > 0$ such that (3.26) is satisfied. The only thing left to check is Condition 5. For $\boldsymbol{\zeta}_i = \sum_{j=1}^{n_i}(y_{ij} - p(\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star))\mathbf{x}_{ij}'$, to prove that for any unit vector $\mathbf{v}_n \in \mathbf{R}^{p_n}$ the sequence $\boldsymbol{\zeta}_i\mathbf{v}_n$ satisfy the Lyapunov condition, we have:

$$\begin{aligned}
E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^3 &= E\left|\sum_{k=1}^{p_n}\mathbf{v}_{nk}\boldsymbol{\zeta}_{ik}\right|^3 \\
&\leq E[\sum_{k=1}^{p_n}\mathbf{v}_{nk}^2\sum_{k=1}^{p_n}\boldsymbol{\zeta}_{ik}^2]^{3/2} \\
&= \left\|\sum_{k=1}^{p_n}\boldsymbol{\zeta}_{ik}^2\right\|_{3/2}^{3/2} \\
&\leq \left(\sum_{k=1}^{p_n}\|\boldsymbol{\zeta}_{ik}^2\|_{3/2}\right)^{3/2} \\
&\leq p_n^{3/2}\max_{1\leq k\leq p_n}E|\boldsymbol{\zeta}_{ik}|^3 \\
&\leq p_n^{3/2}C. \qquad (3.28)
\end{aligned}$$

where C is a constant, since $\boldsymbol{\zeta}_{ik}$ has uniformly bounded $(4r)^{th}$ moment for all $1 \leq k \leq p_n$. On the other hand,

$$\begin{aligned}
E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^2 &= E\left(\sum_{j=1}^{n_i}(y_{ij} - p(\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star))(\mathbf{x}_{ij}\mathbf{v}_n)\right)^2 \\
&= E\left[E\left[\left(\sum_{j=1}^{n_i}(y_{ij} - p(\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star))(\mathbf{x}_{ij}\mathbf{v}_n)\right)^2 \middle| u_i, n_i, \mathbf{x}_{ij}\right]\right] \\
&\geq E\left[E\left[\left(\sum_{j=1}^{n_i}(y_{ij} - p(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u_i))(\mathbf{x}_{ij}\mathbf{v}_n)\right)^2 \middle| u_i, n_i, \mathbf{x}_{ij}\right]\right] \\
&= E\left[E\left[\left(\sum_{j=1}^{n_i} d(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u_i)(\mathbf{x}_{ij}\mathbf{v}_n)\right)^2 \middle| u_i, n_i, \mathbf{x}_{ij}\right]\right] \\
&\geq \frac{1}{2}E\left[d(\mathbf{x}_{ij}\boldsymbol{\beta}_0)(\mathbf{x}_{ij}\mathbf{v})^2\right]
\end{aligned} \tag{3.29}$$

Use the fact that

$$E(\mathbf{x}_{ij}\mathbf{v}_n)^2 = E[\mathbf{v}_n'\mathbf{x}_{ij}'\mathbf{x}_{ij}\mathbf{v}_n] = \mathbf{v}_n'\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\mathbf{v}_n$$

and Assumption 1.4, we have

$$m^* \leq E(\mathbf{x}_{ij}\mathbf{v}_n)^2 \leq M^*.$$

For any a_*,

$$\begin{aligned}
m^* &\leq E(\mathbf{x}_{ij}\mathbf{v}_n)^2 \\
&= E\left[(\mathbf{x}_{ij}\mathbf{v}_n)I[|\mathbf{x}_{ij}\mathbf{v}_n| \leq a_*]\right] + E\left[(\mathbf{x}_{ij}\mathbf{v}_n)I[|\mathbf{x}_{ij}\mathbf{v}_n| \geq a_*]\right] \\
&\leq a_*M^* + E\left[(\mathbf{x}_{ij}\mathbf{v}_n)I[|\mathbf{x}_{ij}\mathbf{v}_n| \geq a_*]\right] \\
&\overset{Holder}{\leq} a_*M^* + \left(E|\mathbf{x}_{ij}\mathbf{v}_n|^{2+\varepsilon}\right)^{\frac{1}{2+\varepsilon}}\left(E\left[I[|\mathbf{x}_{ij}\mathbf{v}_n| \geq a_*]\right]^{\frac{2+\varepsilon}{1+\varepsilon}}\right)^{\frac{1+\varepsilon}{2+\varepsilon}} \\
&\leq a_*M^* + C_\varepsilon^{\frac{1}{2+\varepsilon}} P\left[|\mathbf{x}_{ij}\mathbf{v}_n| \geq a_*\right]^{\frac{1+\varepsilon}{2+\varepsilon}}
\end{aligned} \tag{3.30}$$

For a_* small enough, $m^* - a_* M^* > 0$ and

$$P\left[|\mathbf{x}_{ij}\mathbf{v}_n| \geq a_*\right] \geq \left[\frac{m^* - a_* M^*}{C_\varepsilon^{\frac{1}{2+\varepsilon}}}\right]^{\frac{2+\varepsilon}{1+\varepsilon}} \tag{3.31}$$

On the other hand, for any $A > 0$,

$$P\left[|\mathbf{x}_{ij}\mathbf{v}_n| > a_*, |\mathbf{x}_{ij}\boldsymbol{\beta}_0| > A\right] \leq \frac{E(\mathbf{x}_{ij}\boldsymbol{\beta}_0)^2}{A^2} \leq \frac{M^*\|\boldsymbol{\beta}_0\|^2}{A^2} \tag{3.32}$$

By (3.31) and (3.32),

$$\begin{aligned} P\left[|\mathbf{x}_{ij}\mathbf{v}_n| > a_*, |\mathbf{x}_{ij}\boldsymbol{\beta}_0| \leq A\right] &= P\left[|\mathbf{x}_{ij}\mathbf{v}_n| > a_*\right] - P[[|\mathbf{x}_{ij}\mathbf{v}_n| > a_*, |\mathbf{x}_{ij}\boldsymbol{\beta}_0| > A] \\ &\geq \left[\frac{m^* - a_* M^*}{C_\varepsilon^{\frac{1}{2+\varepsilon}}}\right]^{\frac{2+\varepsilon}{1+\varepsilon}} - \frac{M^*\|\boldsymbol{\beta}_0\|^2}{A^2} \\ &> 0 \end{aligned} \tag{3.33}$$

if we choose a_* to be small enough and A to be large enough.

Therefore,

$$E\left[d(\mathbf{x}_{ij}\boldsymbol{\beta}_0)(\mathbf{x}_{ij}\mathbf{v}_n)^2\right] \geq a_*^2 d(A) P\left[|\mathbf{x}_{ij}\mathbf{v}_n| > a, |\mathbf{x}_{ij}\boldsymbol{\beta}_0| \leq A\right] \geq A_1 > 0,$$

and

$$\sum_{i=1}^{m} \frac{E|\mathbf{v}_n'\boldsymbol{\zeta}_i|^3}{\sigma_{n,\mathbf{v}_n}^3} \leq \frac{m p_n^{3/2} C}{m^{3/2} A_1^{3/2}} \to 0.$$

The last condition in Theorem 3.2 is satisfied. At last, the r has to be the largest of $2\theta/(1-2\theta)$, $\theta/(1-4\theta)$ and $1/[2(1-3\theta)]$, which is the last one when $\theta < 1/4$. □

3.3.4 Limiting Case: $\sigma_u \to 0$

When σ_u is small enough, we expect $\boldsymbol{\beta}_n^\star$ not to be very far from $\boldsymbol{\beta}_0$. Using Taylor expansion, we can approximate the difference $(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0)$ when $\sigma_u = 0^+$:

Theorem 3.5 *When σ_u is in a sufficiently small neighborhood of* 0, *the difference* $(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0)$ *satisfies the equation*

$$\mathbf{G}^*(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0) = \frac{\sigma_u^2}{2}\mathbf{h}^* + O(\sigma_u^4), \tag{3.34}$$

where

$$\mathbf{G}^* = E_{\mathbf{x}}\left[\frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^2}\mathbf{x}'\mathbf{x}\right],$$

and

$$\mathbf{h}^* = E_{\mathbf{x}}\left[\frac{e^{\mathbf{x}\boldsymbol{\beta}_0}(1-e^{\mathbf{x}\boldsymbol{\beta}_0})}{(1+e^{\boldsymbol{\beta}_\mathbf{0}})^3}\mathbf{x}'\right].$$

Lemma 3.1 *When* $\sigma_u = 0$, $\boldsymbol{\beta}_n^\star = \boldsymbol{\beta}_0$.

Proof of Lemma: As the unique solution to $E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n}(\boldsymbol{\beta}_n) = \mathbf{0}]$, $\boldsymbol{\beta}_n^\star \equiv \boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \sigma_u^2)$ is evidently a function of σ_u and $\boldsymbol{\beta}_0$. Let

$$h_*(\sigma_u) \equiv E_u(g^{-1}(\mathbf{x}\boldsymbol{\beta}_0 + u)) = \int \frac{e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_0 z}}\phi(z)dz \tag{3.35}$$

and let

$$g_*(\sigma_u^2) \equiv g^{-1}(\mathbf{x}\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \sigma_u^2)) = \frac{e^{\mathbf{x}\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0,\sigma_u^2)}}{1+e^{\mathbf{x}\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0,\sigma_u^2)}}. \tag{3.36}$$

Then $\boldsymbol{\beta}_n^\star$ satisfies

$$E_{\mathbf{x}}\left[\left(h_*(\sigma_u) - g_*(\sigma_u^2)\right)\mathbf{x}'\right] = \mathbf{0}. \tag{3.37}$$

Moreover, at $\sigma_u = 0$, the model is not misspecified,

$$h_*(0) = g^{-1}(\mathbf{x}\boldsymbol{\beta}_0),$$

and $\boldsymbol{\beta}_n^\star$ satisfies

$$E_{\mathbf{x}}\left[\left(g^{-1}(\mathbf{x}\boldsymbol{\beta}_0) - g^{-1}(\mathbf{x}\boldsymbol{\beta}_n^\star)\right)\mathbf{x}'\right] = \mathbf{0}.$$

The solution of $E[\nabla_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] = \mathbf{0}$ is $\boldsymbol{\beta}_n^\star = \boldsymbol{\beta}_0$, and again it is the only solution by concavity of the log likelihood. □

Remark: Theorem 3.3 and Lemma 3.1 concludes that when all the assumptions are satisfied, at $\sigma_u = 0$, a unique solution to the likelihood equations exists in a neighborhood about $\boldsymbol{\beta}_0$ with probability going to one, in other words, the MLE is consistent. □

Proof of the Theorem: If we write $\boldsymbol{\beta}_n^\star$ as a function of $\boldsymbol{\beta}_0$ and σ_u, from Lemma 3.1 we have

$$\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, 0) = \boldsymbol{\beta}_0.$$

By the Inverse Function Theorem, $\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \sigma_u^2)$ is a smooth function of σ_u^2. So the Taylor expansion of $\boldsymbol{\beta}_n^\star$ at $\sigma_u = 0$ is

$$\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \sigma_u^2) = \boldsymbol{\beta}_0 + \sigma_u^2 \frac{\partial \boldsymbol{\beta}_n^\star}{\partial \sigma_u^2}\Big|_{\sigma_u=0} + O(\sigma_u^4),$$

or

$$\frac{\partial \boldsymbol{\beta}_n^\star}{\partial \sigma_u^2}\Big|_{\sigma_u=0} = \frac{\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0}{\sigma_u^2} + O(\sigma_u^2). \tag{3.38}$$

where the constant in $O(\sigma_u^2)$ is independent of $\mathbf{x}$. The following are true for the function $h_*(\sigma_u)$ defined in (3.35):

$$h_*(0) = \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1 + e^{\mathbf{x}\boldsymbol{\beta}_0}},$$

$$\frac{\partial h_*(0)}{\partial \sigma_u} = \int z \frac{e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}}{(1 + e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z})^2} \phi(z) dz\Big|_{\sigma_u=0} = 0,$$

and

$$\frac{\partial^2 h_*(0)}{\partial \sigma_u^2} = \int z^2 \frac{e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}(1 - e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z})}{(1 + e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z})^3} \phi(z) dz\Big|_{\sigma_u=0} = \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}(1 - e^{\mathbf{x}\boldsymbol{\beta}_0})}{(1 + e^{\mathbf{x}\boldsymbol{\beta}_0})^3}.$$

It is easy to see that $h_*(\sigma_u)$ is an even function of σ_u. Therefore, the Taylor expansion around 0 for $h_*(\sigma_u)$, up to third order, with remainder, is:

$$\begin{aligned} h_*(\sigma_u) &= h_*(0) + \sigma_u \frac{\partial h_*(0)}{\partial \sigma_u} + \frac{\sigma_u^2}{2} \frac{\partial^2 h_*(0)}{\partial \sigma_u^2} + O(\sigma_u^4) \\ &= \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0}} + \frac{\sigma_u^2 e^{\mathbf{x}\boldsymbol{\beta}_0}(1-e^{\mathbf{x}\boldsymbol{\beta}_0})}{2(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^3} + O(\sigma_u^4) \end{aligned} \tag{3.39}$$

where the constant in $O(\sigma_u^4)$ is uniformly bounded and therefore independent of $\mathbf{x}$. On the other hand, the following are true for the function $g_*(\sigma_u^2)$ defined in (3.36):

$$g_*(0) = g^{-1}(\mathbf{x}\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, 0)) = \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0}},$$

$$\begin{aligned} \frac{\partial g_*(0)}{\partial \sigma_u^2} &= \frac{\partial g^{-1}(\mathbf{x}\boldsymbol{\beta}_n^\star)}{\partial \sigma_u^2}\Big|_{\sigma_u=0} = \frac{\partial g^{-1}(\mathbf{x}\boldsymbol{\beta}_n^\star)}{\partial (\mathbf{x}\boldsymbol{\beta}_n^\star)} \cdot \frac{\partial (\mathbf{x}\boldsymbol{\beta}_n^\star)}{\partial \sigma_u^2}\Big|_{\sigma_u=0} \\ &= \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^2} \mathbf{x} \frac{\partial \boldsymbol{\beta}_n^\star}{\partial \sigma_u^2}\Big|_{\sigma_u=0}. \end{aligned} \tag{3.40}$$

Therefore, the Taylor expansion for function $g_*(\sigma_u^2)$ around 0 is

$$\begin{aligned} g_*(\sigma_u^2) &= g_*(0) + \sigma_u^2 \frac{\partial g_*(0)}{\partial \sigma_u^2} + O(\sigma_u^4) \\ &= \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0}} + \sigma_u^2 \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^2} \mathbf{x} \frac{\partial \boldsymbol{\beta}_n^\star}{\partial \sigma_u^2}\Big|_{\sigma_u=0} \\ &= \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0}} + \sigma_u^2 \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^2} \mathbf{x} \left(\frac{\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0}{\sigma_u^2}\right) + O(\sigma_u^4) \\ &= \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0}} + \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^2} \mathbf{x}(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0) + O(\sigma_u^4) \end{aligned} \tag{3.41}$$

where the constant in $O(\sigma_u^4)$ is uniformly bounded for $\mathbf{x}$. Now (3.22) is equal to

$$\begin{aligned} E[G_n(\boldsymbol{\beta}_n^\star)] &= E_{\mathbf{x}}\left[\left(h_*(\sigma_u) - g_*(\sigma_u^2)\right)\mathbf{x}'\right] \\ &= E_{\mathbf{x}}\left[\left(\frac{\sigma_u^2 e^{\mathbf{x}\boldsymbol{\beta}_0}(1-e^{\mathbf{x}\boldsymbol{\beta}_0})}{2(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^3} - \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{(1+e^{\mathbf{x}\boldsymbol{\beta}_0})^2}\mathbf{x}(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0) + O(\sigma_u^4)\right)\mathbf{x}'\right] \\ &= \left[\frac{\sigma_u^2}{2}\mathbf{h}^* - \mathbf{G}^*(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0)\right] + O(\sigma_u^4), \end{aligned}$$

which implies that the root $\boldsymbol{\beta}_n^\star$ satisfies (3.34). □

Remark: The difference $(\boldsymbol{\beta}_* - \boldsymbol{\beta}_0)$ is of the order σ_u^2. Therefore, when σ_u is close to zero, the difference is very small, indicating that the later entries of $\boldsymbol{\beta}_*$ are close to zero. □

3.3.5 Limiting Case: $\sigma_u \to \infty$

The limiting case where $\sigma_u \to \infty$ might not be as realistic as the limiting case of small σ_u, but as a continuous (and infinitely differentiable) function of σ_u, the behavior of $\boldsymbol{\beta}_*$ at large σ_u values should also be studied carefully.

Lemma 3.2 $\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \infty) = \lim_{\sigma_u \to \infty} \boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \sigma_u^2) = \mathbf{0}$.

Proof: First let us look at the function $h_*(\sigma_u)$ defined in (3.35):

$$\lim_{\sigma_u \to \infty} \frac{e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}}{1 + e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}} = \begin{cases} 0 & : \quad z < 0 \\ \frac{e^{\mathbf{x}\boldsymbol{\beta}_0}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0}} & : \quad z = 0 \\ 1 & : \quad z > 0 \end{cases}.$$

Therefore, by Dominated Convergence,

$$h_*(\infty) = \int \lim_{\sigma_u \to \infty} \frac{e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}}{1 + e^{\mathbf{x}\boldsymbol{\beta}_0 + \sigma_u z}} \phi(z) dz = \int_{z>0} \phi(z) dz = \frac{1}{2}.$$

So $\boldsymbol{\beta}_n^\star$ satisfies

$$E_{\mathbf{x}} \left[\left(\frac{1}{2} - \lim_{\sigma_u \to \infty} g_*(\sigma_u) \right) \mathbf{x}' \right] = \mathbf{0}. \tag{3.42}$$

Again, one obvious solution to the above equation is

$$g_*(\infty) = \lim_{\sigma_u \to \infty} g_*(\sigma_u^2) = \lim_{\sigma_u \to \infty} \frac{e^{\mathbf{x}\boldsymbol{\beta}_*(\boldsymbol{\beta}_0, \sigma_u^2)}}{1 + e^{\mathbf{x}\boldsymbol{\beta}_*(\boldsymbol{\beta}_0, \sigma_u^2)}} = \frac{1}{2},$$

or equivalently,

$$\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \infty) = \lim_{\sigma_u \to \infty} \boldsymbol{\beta}^\star(\boldsymbol{\beta}_0, \sigma_u^2) = \mathbf{0}.$$

Suppose there is another solution to (3.42), say, $\tilde{\boldsymbol{\beta}}_n \neq \mathbf{0}$, then

$$E_{\mathbf{x}}\left[\left(\frac{1}{2} - p(\mathbf{x}\tilde{\boldsymbol{\beta}}_n)\right)\mathbf{x}'\right]\tilde{\boldsymbol{\beta}}_n = 0$$

where $p(x) = e^x/(1+e^x)$. Since $(1/2 - p(x))$ and x always have opposite signs, this implies that the random variable $(\mathbf{x}\tilde{\boldsymbol{\beta}}_n)$ degenerates to zero, which is a contradiction to Assumption 1.4 when $\tilde{\boldsymbol{\beta}}_n \neq \mathbf{0}$. Therefore $\boldsymbol{\beta}_n^\star = \mathbf{0}$ is the only solution. □

To see how fast $\boldsymbol{\beta}_*$ is approaching to $\mathbf{0}$ when σ_u approaches infinity, we first need to see how rapidly $h_*(\sigma_u)$ approaches $1/2$ when $\sigma_u \to \infty$.

First, consider the following lemma:

Lemma 3.3 *For large* σ_u,

$$\int_{z>0} e^{-(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \sigma_u z)}\phi(z)dz = e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}\left(\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2\sigma_u^3}\right) + o(\sigma_u^{-3}),$$

and

$$\int_{z<0} e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \sigma_u z}\phi(z)dz = e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}\left(\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2\sigma_u^3}\right) + o(\sigma_u^{-3}).$$

Proof: By Proposition B.2,

$$\begin{aligned}
\int_{z>0} e^{-(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \sigma_u z)}\phi(z)dz &= e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\{-\sigma_u z - \frac{z^2}{2}\}dz \\
&= \exp\{-\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \frac{\sigma_u^2}{2}\}\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\{-\frac{(z+\sigma_u)^2}{2}\}dz \\
&= \exp\{-\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \frac{\sigma_u^2}{2}\}(1 - \Phi(\sigma_u)) \\
&= \exp\{-\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \frac{\sigma_u^2}{2}\}\frac{e^{-\frac{\sigma_u^2}{2}}}{\sigma_u}\left(\frac{1}{\sqrt{2\pi}} - \frac{1}{2\sqrt{\sigma_u^2 + \sigma_u^4}}\right) + o(\sigma_u^{-3})
\end{aligned}$$

$$= \frac{e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}}{\sigma_u}\left(\frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma_u^2\sqrt{1+\sigma_u^{-2}}}\right) + o(\sigma^{-3})$$
$$= e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}\left(\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2\sigma_u^3}\right) + o(\sigma_u^{-3}). \quad (3.43)$$

And similarly, by change of variable we get

$$\int_{z<0} e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0+\sigma_u z}\phi(z)dz = e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}\exp\{\sigma_u z - \frac{z^2}{2}\}dz$$
$$= e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}\left(\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2\sigma_u^3}\right) + o(\sigma_u^{-3}). \quad (3.44)$$

□

Remark: It is worth mentioning that the constants in $o(\sigma_u^{-3})$ in (3.43) and (3.44) involve $e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}$ and $e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}$, respectively.

The following lemma states the rate at which $h_*(\sigma_u)$ goes to 1/2 when $\sigma_u \to \infty$.

Lemma 3.4 *For large σ_u,*

$$h_*(\sigma_u) - \frac{1}{2} = C^{(1)}\sigma_u^{-1} + C^{(2)}\sigma_u^{-3} + o(\sigma_u^{-3})$$

where

$$C^{(1)} = \frac{e^{\mathbf{x}\boldsymbol{\beta}_0} - e^{-\mathbf{x}\boldsymbol{\beta}_0}}{\sqrt{2\pi}},$$

and

$$C^{(2)} = \frac{e^{-\mathbf{x}\boldsymbol{\beta}_0} - e^{\mathbf{x}\boldsymbol{\beta}_0}}{2}.$$

Proof: When σ_u is large and $z < 0$,

$$\frac{e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}} \sim e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z},$$

and for $z > 0$,

$$\frac{e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}} \approx 1 - e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0-\sigma_u z}.$$

Therefore,

$$\begin{aligned} h_*(\sigma_u) - \frac{1}{2} &= \int \frac{e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}}{1+e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}}\phi(z)dz - \frac{1}{2} \\ &\approx \int_{z<0} e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}\phi(z)dz + \int_{z>0}[1-e^{-\mathbf{x}\boldsymbol{\beta}_0-\sigma_u z}]\phi(z)dz - \Phi(0) \\ &= \int_{z<0} e^{\mathbf{x}\boldsymbol{\beta}_0+\sigma_u z}\phi(z)dz - \int_{z>0} e^{-\mathbf{x}\boldsymbol{\beta}_0-\sigma_u z}\phi(z)dz \\ &\approx (e^{\mathbf{x}\boldsymbol{\beta}_0} - e^{-\mathbf{x}\boldsymbol{\beta}_0})\left(\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2\sigma_u^3}\right) \\ &\equiv C^{(1)}\sigma_u^{-1} + C^{(2)}\sigma_u^{-3} + o(\sigma_u^{-3}). \end{aligned} \tag{3.45}$$

□

Remark: The constant in $o(\sigma_u^{-3})$ in (3.45) involves both $e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}$ and $e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}$.

□

If in addition we assume that

Assumption 3.3 *The expectation* $E[e^{|\mathbf{x}\boldsymbol{\beta}_0|}]$ *exists, and*

$$\boldsymbol{\mu}_e \equiv E_{\mathbf{x}}\left[\left(e^{\mathbf{x}\boldsymbol{\beta}_0} - e^{-\mathbf{x}\boldsymbol{\beta}_0}\right)\mathbf{x}'\right].$$

Then when taking expectation with respect to the $o(\sigma_u^{-3})$ terms the constants involving $e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}$ or $e^{-\mathbf{x}_{ij}\boldsymbol{\beta}_0}$ are all bounded and $o(\sigma_u^{-3})$ can be taken out of the expectation. Therefore we have the following Theorem:

Theorem 3.6 *The solution to (3.22) when* σ_u *is large, is*

$$\boldsymbol{\beta}_n^\star = \frac{4}{\sigma_u}\left[\frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma_u^2}\right]\left[\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\right]^{-1}\boldsymbol{\mu}_e + o(\sigma_u^{-3})E_{\mathbf{x}}[\mathbf{x}'] + o(\|\boldsymbol{\beta}_n^\star\|^2). \tag{3.46}$$

Proof: Note that the Taylor expansion for the function

$$h(t) = g^{-1}(t) = \frac{e^t}{1+e^t}$$

around zero is

$$h(t) = h(0) + h'(0)t + h''(0)\frac{t^2}{2} + o(t^2) = \frac{1}{2} + \frac{1}{4}t + O(t^2).$$

Therefore the Taylor expansion for $g_*(\sigma_u^2) = h(\mathbf{x}\boldsymbol{\beta}_n^\star)$ around $\boldsymbol{\beta}_n^\star = \mathbf{0}$ is

$$g_*(\sigma_u^2) = \frac{1}{2} + \frac{1}{4}\mathbf{x}\boldsymbol{\beta}_n^\star(\boldsymbol{\beta}_0, \sigma_u^2) + o((\mathbf{x}\boldsymbol{\beta}_n^\star)^2).$$

As $\boldsymbol{\beta}_n^\star$ gets close to $\mathbf{0}$, $\mathbf{x}\boldsymbol{\beta}_n^\star$ will get close to 0 as well, and when taken the expectation with respect to $\mathbf{x}$, the $o((\mathbf{x}\boldsymbol{\beta}_n^\star)^2)$ becomes $o(\|\boldsymbol{\beta}_n^\star\|^2)$. Using this and Lemma 3.4 in (3.22), we find

$$\begin{aligned}
& E_{\mathbf{x}}\left[\left(h_*(\sigma_u) - g_*(\sigma_u^2)\right)\mathbf{x}'\right] \\
= \; & E_{\mathbf{x}}\left[\left(\frac{1}{2} + C^{(1)}\sigma_u^{-1} + C^{(2)}\sigma_u^{-3} + o(\sigma_u^{-3}) - \frac{1}{2} - \frac{1}{4}\mathbf{x}\boldsymbol{\beta}_n^\star + o((\mathbf{x}\boldsymbol{\beta}_n^\star)^2)\right)\mathbf{x}'\right] \\
= \; & E_{\mathbf{x}}\left[\left(C^{(1)}\sigma_u^{-1} + C^{(2)}\sigma_u^{-3} + o(\sigma_u^{-3})\right)\mathbf{x}'\right] - \frac{1}{4}\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\boldsymbol{\beta}_n^\star + o(\|\boldsymbol{\beta}_n^\star\|^2) \\
= \; & \left[\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2\sigma_u^3}\right]\boldsymbol{\mu}_e + o(\sigma_u^{-3})E_{\mathbf{x}}[\mathbf{x}'] - \frac{1}{4}\boldsymbol{\Sigma}_{\mathbf{x}}^{(n)}\boldsymbol{\beta}_n^\star + o(\|\boldsymbol{\beta}_n^\star\|^2).
\end{aligned}$$

which means that $\boldsymbol{\beta}_n^\star$ satisfies (3.46). □

Remark: It is obvious that with σ_u^2 large, the effect of $\mathbf{x}_{ij}$ are "washed out" and all the entries of $\boldsymbol{\beta}_n^\star$ are close to zero.

3.4 Poisson Regression

In this part of the discussion, we impose Assumptions 1.2-1.4 as well as Assumption 3.1. In the Poisson regression model, the log likelihood function of the model with sample size n is

$$l_n(\boldsymbol{\beta}_n) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left(\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!}\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[-\mu_{ij} + y_{ij}\log\mu_{ij} - \log(y_{ij}!)\right], \tag{3.47}$$

where

$$\mu_{ij} = Ey_{ij}$$

with the expectation taken under the working model. In the GLM notations,

$$\tau = 1,$$

$$\gamma_{ij} = \log\mu_{ij},$$

$$b(\gamma_{ij}) = \mu_{ij} = e^{\gamma_{ij}}$$

and

$$c(y_{ij}, \tau) = \log(y_{ij}!).$$

With the canonical link

$$g(\mu) = \log(\mu),$$

(3.47) becomes

$$l_n(\boldsymbol{\beta}_n) = \sum_i\sum_j\left[y_{ij}\mathbf{x}_{ij}\boldsymbol{\beta}_n - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n} - \log(y_{ij}!)\right].$$

Without specifying the distribution f_u, we can see that

$$\frac{\partial l_n(\boldsymbol{\beta}_n)}{\partial\boldsymbol{\beta}_n} = \sum_i\sum_j\left[y_{ij} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}\right]\mathbf{x}'_{ij}.$$

Therefore,

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n) = \frac{1}{m}\frac{\partial l_n(\boldsymbol{\beta}_n)}{\partial\boldsymbol{\beta}_n} = \frac{1}{m}\sum_i\sum_j\left[y_{ij} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}\right]\mathbf{x}'_{ij} \tag{3.48}$$

and

$$E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] = E_{\mathbf{x}}\left[\left(E_u(g^{-1}(\mathbf{x}\boldsymbol{\beta}_0 + u)) - e^{\mathbf{x}\boldsymbol{\beta}_n}\right)\mathbf{x}'\right]$$

where

$$E_u(g^{-1}(\mathbf{x}\boldsymbol{\beta}_0 + u)) = \int e^{\mathbf{x}\boldsymbol{\beta}_0 + u} f_U(u)du.$$

The solution $\boldsymbol{\beta}_n^\star$ to the equations $E[\nabla_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] = \mathbf{0}$ therefore satisfies

$$E_{\mathbf{x}}\left[\left(\int e^{\mathbf{x}\boldsymbol{\beta}_0 + u} f_U(u)du - e^{\mathbf{x}\boldsymbol{\beta}_n^\star}\right)\mathbf{x}'\right] = \mathbf{0}, \tag{3.49}$$

and it is unique because of concavity of the log likelihood.

Furthermore, we have

Theorem 3.7 *Under the Assumptions 1.2-1.4, as $n \to \infty$, a unique solution $\hat{\boldsymbol{\beta}}_n$ to $G_n(\boldsymbol{\beta}_n) = \mathbf{0}$ exists in a neighborhood about $\boldsymbol{\beta}_n^\star$ with probability going to one, and $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^\star\| = O_p(\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\|)$.*

Proof: We can use the same arguments as in the proof of Theorem 3.3. Note that on a compact set of $\boldsymbol{\beta}_n$ the function $e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n}$ is bounded. Assuming $\boldsymbol{\beta}_n^\star$ exists, then $-\boldsymbol{\nabla}^{\otimes 2}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)$ is bounded in the matrices sense by a constant multiplied by $\mathbf{I}_{p_n}$, and $\|\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n^\star)\| \leq O_p(p_n/\sqrt{n})$ by Theorem 1.1 and the boundedness of $(y_{ij} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_n^\star})$ uniformly over i and j. □

Theorem 3.8 *Under conditions and assumptions in Theorem 3.4, the $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n^\star)$ is asymptotically normal.*

Proof: Note that in the proof of Theorem 3.4 we only need $d(x)$ to be locally uniformly positive. The function e^x has the same property and we can follow exactly same arguments to prove the asymptotic normality of $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$. Therefore details of the proof will not be supplied here. □

Since the Poisson model assumes that the mean and variance of the data are the same, there are many discussions in the literature concerning the validity of this assumption, in particular, when the variance of the data is bigger than what the model explains(in Poisson, the mean), which is often called "overdispersion". Papers by Breslow [3] [4], Dean and Lawless [7] and Wilson [33] have considered overdispersion relative to Poisson regression and log-linear models. In these discussions the authors acknowledged the fact that the variability in the data can not be fully explained by the mean-variance relationship assumed in a Poisson regression. Different test statistics were proposed to test overdispersion in Poisson model, or Quasi-Likelihood Equations are solved instead of log-likelihood equations to get a better variance estimator for the coefficients $\boldsymbol{\beta}$. Clearly $\boldsymbol{\beta}_n^\star$ depends on the distribution of u. We will discuss the value of $\boldsymbol{\beta}_n^\star$ with two different assumptions on the conditional mean μ_{ij}. One is the Normal random intercept: $u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$, and the other is $u_i \overset{iid}{\sim}$ Gamma(α, β). Whenever $\boldsymbol{\beta}_n^\star$ exists, it is unique because of concavity of the loglikelihood for Poisson model.

3.4.1 Normal Random Effects

Under Assumption 3.2,

$$\begin{aligned} E_u(g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u)) &= \int e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \sigma_u z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \exp\{\mathbf{x}_{ij}\boldsymbol{\beta}_0 + \frac{\sigma_u^2}{2}\}. \end{aligned}$$

Therefore

$$E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] = E_{\mathbf{x}}\left[\left(e^{\mathbf{x}\boldsymbol{\beta}_0 + \frac{\sigma_u^2}{2}} - e^{\mathbf{x}\boldsymbol{\beta}_n}\right)\mathbf{x}'\right]. \tag{3.50}$$

One obvious solution to (3.50) is

$$\boldsymbol{\beta}_n^\star = \boldsymbol{\beta}_0 + \sigma_u^2/2,$$

where a scalar $\sigma_u^2/2$ is added to the first element of $\boldsymbol{\beta}_0$, i.e., the intercept term of $\boldsymbol{\beta}_0$. By the concavity of the log likelihood, it is unique. Therefore, a normal random intercept in Poisson regression places an offset on the intercept estimate. This is true whenever the working model contains an intercept term. Even when the true model does not contain a fixed intercept, the intercept estimate of the working model will converge to $\sigma_u^2/2$. Therefore with Normal random intercept, all the coefficient estimators are consistent except for the intercept. This is a well-know fact mentioned in McCulloch and Searle [18].

3.4.2 Gamma-Poisson Model

We next assume instead of Assumption 3.2 that

Assumption 3.4 $u_i \overset{iid}{\sim} Gamma(\alpha, \beta)$, *for* $0 < \beta < 1$ *and* $\alpha > 0$.

Then

$$\begin{aligned} E_u(g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u)) &= \int_0^\infty e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0 + u} \frac{u^{\alpha-1}e^{-\frac{u}{\beta}}}{\beta^\alpha \Gamma(\alpha)} du \\ &= \frac{e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty e^{-u(\frac{1}{\beta}-1)} u^{\alpha-1} du \\ &= e^{\mathbf{x}_{ij}\boldsymbol{\beta}_0}(1-\beta)^{-\alpha} \end{aligned}$$

and

$$E[\boldsymbol{\nabla}_{\boldsymbol{\beta}_n} G_n(\boldsymbol{\beta}_n)] = E_{\mathbf{x}}\left[\left(e^{\mathbf{x}\boldsymbol{\beta}_0}(1-\beta)^{-\alpha} - e^{\mathbf{x}\boldsymbol{\beta}}\right)\mathbf{x}'\right]. \tag{3.51}$$

Therefore the unique solution to (3.51) is

$$\boldsymbol{\beta}_n^\star = -\alpha \ln(1-\beta)\boldsymbol{\beta}_0. \tag{3.52}$$

$\boldsymbol{\beta}_n^\star$ contains the same number of non-zero entries as $\boldsymbol{\beta}_0$, and with $\alpha > 0$ and $0 < \beta < 1$, $-\alpha \ln(1-\beta) > 0$, which means that every entry of $\boldsymbol{\beta}_n^\star$ has the same sign as that of $\boldsymbol{\beta}_0$. Whether the entries of $\boldsymbol{\beta}_n^\star$ are larger or smaller in absolute value than the corresponding entries of $\boldsymbol{\beta}_0$ depends on the values of α and β.

- $1 - e^{-1/\alpha} < \beta < 1$, then elements of $\boldsymbol{\beta}_*$ are larger in absolute value than those of $\boldsymbol{\beta}_0$;
- $0 < \beta < 1 - e^{-1/\alpha}$, then elements of $\boldsymbol{\beta}_*$ are smaller in absolute value than those of $\boldsymbol{\beta}_0$;
- $\beta = 1 - e^{-1/\alpha}$, $\boldsymbol{\beta}_* = \boldsymbol{\beta}_0$.

It is obvious that at least on the basis of Kullback-Leibler minimization, we should not have a coefficient that is falsely large, since the coefficient estimates are consistent in the sense that when $n \to \infty$, all the zero-entries of the coefficients should have estimates that are close to zero. But again the variance estimator under the working model does not account for the extra variation that's provided by the random effect, and therefore in hypothesis testing, the smaller variance estimator can lead to false inference, picking coefficients that are not really in the model.

3.5 Computations and Simulations

As mentioned before, most ML equations in Generalized Linear Models do not have a closed-form solution. In Logistic-normal regression discussed in Section 3.3, we can approximate $\boldsymbol{\beta}_n^\star$ to top order when σ_u is extremely large or small, but can not do so when σ_u is moderate; we postponed discussion of the case where the link function is misspecified. We will demonstrate the behavior of $\boldsymbol{\beta}_n^\star$ and $\hat{\boldsymbol{\beta}}_n$ with the help of numerical computation and simulation studies.

3.5.1 Logistic Regression: Moderate σ_u

We showed in Section 3.3 in both extreme cases ($\sigma_u \to 0$ and $\sigma_u \to \infty$) that the entries of $\boldsymbol{\beta}_n^\star$ get small for indices larger than $(p^* + 1)$. This means that when σ_u is in extreme ranges, the variables that are not in the true model will not have significant coefficient estimates. We want to know if this is also true when the value of σ_u is in the moderate range. Since Taylor expansion is not an option when σ_u is in the moderate range, we first compute $\boldsymbol{\beta}_n^\star$ numerically and then carry out a simulation to check agreement with the theoretical calculation.

Suppose that the rows of the design matrix $\mathbf{X}_n$ are iid from some distribution $f_{\mathbf{x}}$. The data summarized in the tables were generated from a conditional binomial distribution given $\mathbf{x}$ using iid normal random variable u with mean 0 and variance σ_u^2, according to the true model, then fitted by a logistic regression (the working model). The average cluster size is taken to be discrete uniform with $N_1 = 5$ and in each dataset there are $m = 200$ clusters. Therefore we have datasets of size 1000.

We also tried different choices of the distribution $f_{\mathbf{x}}$, with the rows of $\mathbf{X}_n$ iid multivariate normal with different choices of the variance-covariance matrix $\mathbf{\Sigma}_{\mathbf{x}}$ (in order to control the correlation between the rows of $\mathbf{X}_n$ to see if high correlation among variables would give $\boldsymbol{\beta}_n^\star$ very different from the low-correlation case. We also tried discrete distributions of $\mathbf{x}$ where we could also control the correlations among variables via the definition of the joint density function. To see whether "added" variables (columns $p^* + 1, \cdots, p_n$ of $\mathbf{X}_n$) highly correlated with the earlier variables would have coefficient estimates that are very different from those uncorrelated (or not highly correlated) with the p^* true variables, we also arranged different variables to be added into the model. What we found in all of these cases was vary similar. We display results only for the case of binary $\mathbf{X}_n$ entries as a demonstration.

In this particular example, $\mathbf{X}_n$ has binary rows. There are four true effects in the model: three binary random variables, and the interaction of two of them. The added variables include another variable that is independent of them, and the remaining two interaction terms of the variables. Table 3.1 illustrates $\boldsymbol{\beta}_{1000}^\star$ when σ_u is in different ranges, and Table 3.2 displays the average coefficient estimates $\bar{\hat{\boldsymbol{\beta}}}_{1000}$ in a simulation of 1000 repetitions of datasets of size 1000.

We can see from Table 3.1 that for the zero elements of $\boldsymbol{\beta}_0$, the corresponding coefficient estimates are also close to zero, throughout the range of σ_u; for the non-zero elements of $\boldsymbol{\beta}_0$, the corresponding coefficient estimates have the same signs, but are attenuated. The extent to which the coefficient estimates are attenuated is determined by σ_u^2. The bigger σ_u^2 is, the bigger the percentage is. For the same σ_u^2, this ratio is roughly the same across different entries of $\boldsymbol{\beta}_0$, leading us to believe that

Table 3.1: $\boldsymbol{\beta}_n^\star$ at different values of σ_u^2, and the percentage of relative error with respect to $\boldsymbol{\beta}_0$. The first of the two columns for each σ_u^2 value is the numerical value of $\boldsymbol{\beta}_n^\star$, and the second column demonstrates the ratio of $\|\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0\|/\|\boldsymbol{\beta}_0\|$. The later rows have blanks because $\boldsymbol{\beta}_0$ is zero in those rows.

	$\sigma_u^2 = .1$		$\sigma_u^2 = .5$		$\sigma_u^2 = 1$		$\sigma_u^2 = 4$		$\sigma_u^2 = 10$		$\sigma_u^2 = 100$	
$\boldsymbol{\beta}_0$	$\boldsymbol{\beta}_n^\star$	%	$\boldsymbol{\beta}_n^\star$	%	$\boldsymbol{\beta}_n^\star$	%	$\boldsymbol{\beta}_n^\star$	%	$\boldsymbol{\beta}_n^\star$	%	$\boldsymbol{\beta}_n^\star$	%
0.5	0.49	2.4	0.45	10.0	0.41	17.2	0.36	27.2	0.22	56.0	0.03	93.6
0.7	0.69	2.0	0.64	9.0	0.59	16.0	0.52	26.1	0.31	55.1	0.06	91.7
-1	-0.98	2.3	-0.90	10.0	-0.83	17.2	-0.73	27.2	-0.44	56.1	-0.06	93.7
0.4	0.39	1.8	0.37	8.5	0.34	15.5	0.30	25.5	0.18	55.3	0.04	90.8
0.6	0.58	2.7	0.54	10.8	0.49	18.2	0.43	28.2	0.26	56.7	0.03	95.3
0	0.00		0.00		0.00		0.00		0.00		0.01	
0	0.00		0.00		0.002		0.00		0.00		-0.04	
0	-0.00		0.00		0.00		0.00		0.00		-0.01	

Table 3.2: $\hat{\boldsymbol{\beta}}_{1000}$ vs $\boldsymbol{\beta}^{\star}_{1000}$: Comparing MLE $\hat{\boldsymbol{\beta}}_n$ under the working model to $\boldsymbol{\beta}^{\star}_n$ when the sample size is large. For each value of σ_u^2, the first column gives the average of $\hat{\boldsymbol{\beta}}_{1000}$ and the second column gives $\boldsymbol{\beta}^{\star}_{1000}$.

	$\sigma_u^2 = .1$		$\sigma_u^2 = .5$		$\sigma_u^2 = 1$		$\sigma_u^2 = 4$		$\sigma_u^2 = 10$	
$\boldsymbol{\beta}_0$	$\hat{\boldsymbol{\beta}}_n$	$\boldsymbol{\beta}^{\star}_n$	$\hat{\boldsymbol{\beta}}_n$	$\boldsymbol{\beta}^{\star}_n$	$\hat{\boldsymbol{\beta}}_n$	$\boldsymbol{\beta}^{\star}_n$	$\hat{\boldsymbol{\beta}}_n$	$\boldsymbol{\beta}^{\star}_n$	$\hat{\boldsymbol{\beta}}_n$	$\boldsymbol{\beta}^{\star}_n$
0.5	0.49	0.49	0.45	0.45	0.41	0.41	0.32	0.36	0.23	0.22
0.7	0.69	0.69	0.64	0.64	0.59	0.59	0.46	0.52	0.31	0.31
-1	-0.99	-0.98	-0.91	-0.90	-0.86	-0.83	-0.61	-0.73	-0.47	-0.44
0.4	0.41	0.39	0.38	0.37	0.37	0.34	0.25	0.30	0.22	0.18
0.6	0.59	0.58	0.54	0.54	0.52	0.49	0.34	0.43	0.33	0.26
0	0.01	0.00	0.00	0.00	0.00	0.00	-0.02	0.00	-0.03	0.00
0	0.00	0.00	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.05	0.00
0	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00

the ignored random effect is the main reason that $\boldsymbol{\beta}_n^\star$ is different from $\boldsymbol{\beta}_0$. Extreme values of σ_u in Table 3.1 (cases where $\sigma_u^2 = .1$ or $\sigma_u^2 = 100.$) also confirm the findings in Section 3.3.

The two columns corresponding to the same σ_u^2 value are quite close when σ_u^2 is small. When σ_u^2 gets to the moderate range $1 \leq \sigma_u^2 \leq 10$, $\boldsymbol{\beta}_n^\star$ is not as well approximated by the average. This could be due to the simulation errors and bigger variance for the data. When σ_u^2 gets too large, the effect of $\mathbf{x}_{ij}$'s are "washed out" by the random effect that has a large variance, the coefficients are close to zero. Theorem 3.3 in Section 3.3 is confirmed by Table 3.2.

As we see in the linear model, omitting the random effect will leave the variance estimators biased. The estimated standard error for the coefficient estimates are always smaller than the actual standard error, as shown in Table 3.3. Let $\hat{\boldsymbol{\beta}}_{1000,i}^{(k)}$ denote the k^{th} element of $\hat{\boldsymbol{\beta}}_{1000}$ at the i^{th} simulation, and diag($\mathbf{M}$) denote the diagonal elements of matrix $\mathbf{M}$; then the k^{th} element of SD_{emp} and SD_{est} are

$$SD_{\text{emp}}^{(k)} = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}\left[\hat{\boldsymbol{\beta}}_{1000,i}^{(k)} - \bar{\hat{\boldsymbol{\beta}}}_{1000}^{(k)}\right]^2},$$

and

$$SD_{\text{est}}^{(k)} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{SD}_i^{(k)}$$

where $\hat{SD}^{(k)} = \sqrt{\text{diag}\left[(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right]_k}$. The weight-matrix $\mathbf{W}$ is defined in (3.10).

SD_{emp} is always bigger than SD_{est}, and the bigger σ_u is, the bigger the difference between SD_{emp} and the corresponding SD_{est} is. This is because in the usual generalized linear models we use the inverse of Fisher information as our asymptotic variance estimator and it is obviously biased when the model is

Table 3.3: Bias in Variance Estimation–Empirical Standard Error vs Estimated Standard Error in A Simulation. For each σ_u value, the column "SD_{emp}" is the empirical standard error in 1000 repetitions, while the column "SD_{est}" is the average of the estimated standard errors in 1000 repetitions.

	$\sigma_u^2 = 0.5$		$\sigma_u^2 = 1$		$\sigma_u^2 = 4$		$\sigma_u^2 = 10$	
$\boldsymbol{\beta}_0$	SD_{emp}	SD_{est}	SD_{emp}	SD_{est}	SD_{emp}	SD_{est}	SD_{emp}	SD_{est}
0.5	0.1903	0.1621	0.2317	0.1730	0.2525	0.1469	0.2486	0.1296
0.7	0.7854	0.6496	0.8576	0.6423	0.9761	0.6185	1.1123	0.5946
-1	0.5290	0.4189	0.4834	0.3594	0.5102	0.3099	0.5554	0.2893
0.4	0.2609	0.2252	0.2847	0.2194	0.3630	0.2090	0.4637	0.2447
0.6	0.6083	0.4892	0.6100	0.4555	0.7063	0.4352	0.7693	0.4006
0	0.8031	0.6670	0.8157	0.6183	0.9126	0.5835	1.2153	0.6363
0	0.6613	0.5466	0.6402	0.4800	0.6989	0.4244	0.9994	0.5164
0	0.2747	0.2252	0.2876	0.2194	0.3291	0.2090	0.4765	0.2447

misspecified.

Also as we have seen in linear models, there is a robust version of variance estimator that converges to the true variance in probability. Table 3.4 compares several estimates for standard deviation at different σ_u levels. The k^{th} element of $\hat{SD}_R$ and SD_R are

$$\hat{SD}_R = \sqrt{\text{diag}\left(\mathbf{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\beta}}_n)\mathbf{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n\right)_k}$$

and

$$SD_R^{(k)} = \sqrt{\text{diag}\left(E[\mathbf{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)\mathbf{B}_n(\hat{\boldsymbol{\beta}}_n)\mathbf{A}_n^{-1}(\hat{\boldsymbol{\beta}}_n)]\right)_k},$$

respectively. Definitions of $\mathbf{A}_n$ and $\mathbf{B}_n$ can be found in (2.30) and (2.31).

$\hat{SD}_R$ is close to SD_{emp} at all levels of σ_u^2. Since SD_{emp} is the closest we get for the estimator of standard deviation of $\hat{\boldsymbol{\beta}}_n$, this suggests that the robust "sandwich" variance estimator is actually doing a good job estimating the true variance of the coefficient estimates. The empirical standard deviation SD_{emp} is usually larger than the other two because this is the true standard deviation of the $\hat{\boldsymbol{\beta}}_n$'s in a repetition of 1000 datasets sharing the same design matrix for the fixed effect. Part of the variability also comes from the sampling variablility in the simulations.

3.5.2 Another Kind of Misspecification

In this section, we no longer assume that the link function $g(\cdot)$ or functions $b(\cdot)$ and $c(\cdot)$ are correctly specified. From (3.13) we can see that only the true link function $g^*(\cdot)$ is involved in solving $\boldsymbol{\beta}_n^\star$. Therefore Theorem 3.3 should still work as long as both $g^*(\cdot)$ and $f_U(\cdot)$ behave well enough so that the conditions in the

Table 3.4: The Variance Estimators: Robust vs Empirical. For each σ_u^2 value the first column is the numerical calculation of the theoretical robust standard deviation estimator(SD_R), the second column is the corresponding average in a simulation run ($\hat{SD}_R$), and the last column is the empirical standard deviation of $\hat{\boldsymbol{\beta}}_n$ in 1000 repetitions.

	$\sigma_u^2 = .5$			$\sigma_u^2 = 1$			$\sigma_u^2 = 4$		
No.	SD_R	$\hat{SD}_R$	SD_{emp}	SD_R	$\hat{SD}_R$	SD_{emp}	SD_R	$\hat{SD}_R$	SD_{emp}
1	0.174	0.185	0.188	0.191	0.201	0.198	0.236	0.255	0.264
2	0.728	0.815	0.849	0.795	0.981	0.999	0.970	1.103	1.129
3	0.388	0.415	0.450	0.429	0.483	0.519	0.528	0.553	0.579
4	0.261	0.347	0.377	0.283	0.389	0.446	0.341	0.406	0.434
5	0.515	0.564	0.591	0.564	0.683	0.699	0.691	0.771	0.809
6	0.728	0.893	0.952	0.794	1.006	1.067	0.967	1.102	1.126
7	0.564	0.667	0.710	0.619	0.778	0.882	0.763	0,856	0.906
8	0.261	0.347	0.380	0.283	0.389	0.437	0.341	0.406	0.421

theorem are satisfied. If we assume that the true model is defined as (3.1), and the working model is a logistic regression model, we wish to see if the results in Section 3.3, i.e. the variables that are not in the true model will not have significantly large nonzero coefficient estimates, are still valid under the wrong link function.

Table 3.5 demonstrates the difference between $\boldsymbol{\beta}_n^\star$ under the right and the wrong link function. In this experiment, we have three discrete variables X_1, X_2 and X_3 that we consider to be the fixed effect, as well as the interaction of X_1 and X_3. To compare the effect of correlation among variables in $\boldsymbol{\beta}_n^\star$, the three added variables we consider in the model are X_4, another variable that has correlation with the three true variables but is not a function of any of them, and two variables that are functions of the true variables, one being the interaction between X_2 and X_3, and the other being the indicator $X_5 = I[X_1 \geq X_2]$. The random intercept of this experiment is assumed to be iid Normal variates with mean 0 and variance σ_u^2.

Let F_{Beta} be the cdf of $Beta(1,1)$, we use the function

$$g^{*-1}(x) = F_{Beta}(\arctan(x)/\pi + 1/2) \tag{3.53}$$

as our true link function. It is easy to see that with g^* defined in (3.53) and f_U the normal density, the conditions in Theore 3.3 are satisfied and the MLE $\hat{\boldsymbol{\beta}}_n$ converges to $\boldsymbol{\beta}_n^\star$ in probability. We numerically calculate the value of $\boldsymbol{\beta}_n^\star$ at four levels of σ_u^2 values, $\sigma_u^2 = 0$, 0.1, 0.5 and 1. Also to compare with the results of Section 3.3, we list the value $\boldsymbol{\beta}_n^\star$ at the corresponding σ_u level when the link function is correctly specified.

From Table 3.5 we can see that with the wrong link function, the behavior of

Table 3.5: Wrong Link vs Right Link: the effect of the true link function g^* on $\boldsymbol{\beta}_n^\star$. The first column lists the variables that are included in the working model, and the second column is the corresponding coefficients of these variables under the true model. Note that the last variables are not in the true model ($\boldsymbol{\beta}_0 = 0$ in the last three entries). For each of the Wrong Link or Right Link column, four levels of σ_u^2 values are considered: $\sigma_u^2 = 0$, 0.1, 0.5, 1. In the columns are the numerically calculated $\boldsymbol{\beta}_n^\star$ values.

		Wrong Link				Right Link			
X	$\boldsymbol{\beta}_0$	0	0.1	0.5	1.0	0	0.1	0.5	1.0
1	0.5	0.674	0.715	0.707	0.606	0.500	0.488	0.446	0.403
X_1	0.4	0.285	0.318	0.367	0.373	0.400	0.397	0.384	0.366
X_2	-0.6	-0.890	-0.890	-0.807	-0.690	-0.600	-0.587	-0.541	-0.495
X_3	0.3	0.443	0.409	0.332	0.297	0.300	0.295	0.280	0.264
X_1X_3	-0.7	-0.890	-0.872	-0.796	-0.772	-0.700	-0.689	-0.650	-0.609
X_4	0.0	-0.028	-0.020	-0.008	-0.002	0.000	0.000	0.002	0.002
X_2X_3	0.0	0.024	0.028	0.027	0.016	0.000	-0.001	-0.004	-0.006
X_5	0.0	0.328	0.207	0.021	0.015	0.000	-0.002	-0.004	-0.002

$\boldsymbol{\beta}_n^\star$ can be quite different from with the right link function. The first thing that draws our attention is the bottom line of Table 3.5, where the variable $X_5 = I[X_1 \geq X_2]$, which is not in the true model, has a nonzero coefficient even when $\sigma_u = 0$. The entry of $\boldsymbol{\beta}_n^\star$ corresponding to X_5 stays nonzero for small σ_u^2 values, and then shrinks to much smaller value when σ_u^2 gets larger. This is significant because it is different from what we have seen in Section 3.3 or Table 3.1: We actually find a situation where a variable that is not in the model has a significant nonzero coefficient, and will be falsely included in the model.

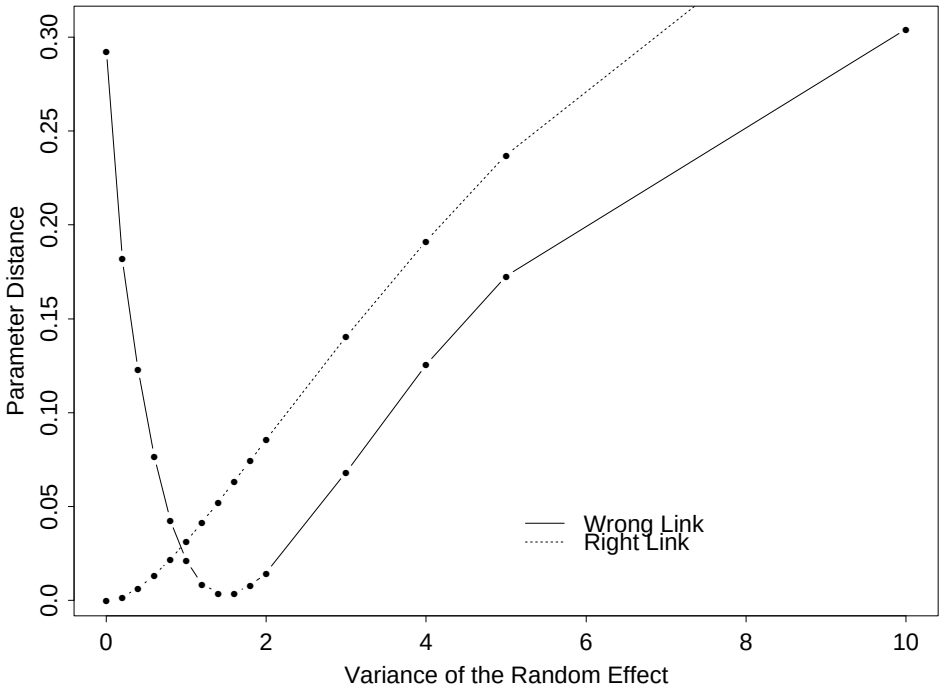

Figure 3.1: The distance between $\boldsymbol{\beta}_n^\star$ and $\boldsymbol{\beta}_0$ ($\|\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0\|^2$) when the link function is correctly specified (dotted line) and when the link function is incorrectly specified (solid line).

As seen in Table 3.1, when σ_u is small and the link function is right, the difference between $\boldsymbol{\beta}_n^\star$ and $\boldsymbol{\beta}_0$, i.e., the Euclidean norm of the vector $(\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0)$, is small, and the bigger σ_u^2 is, the bigger the difference is. In the misspecified link case, we see different behavior. The difference between $\boldsymbol{\beta}_n^\star$ and $\boldsymbol{\beta}_0$ is big when σ_u^2 is small and it gets smaller when σ_u^2 gets larger. It grows large again at larger σ_u^2 values, and when σ_u^2 is extremely large, $\boldsymbol{\beta}_n^\star$ with the wrong link function is essentially the same as that with the right link function–they tend to be close to zero. Figure 3.1 shows this effect by drawing the distance $\|\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0\|^2$ as a function of σ_u^2. When the link function is correct, the distance $\|\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0\|^2$ is an increasing function of σ_u^2, going

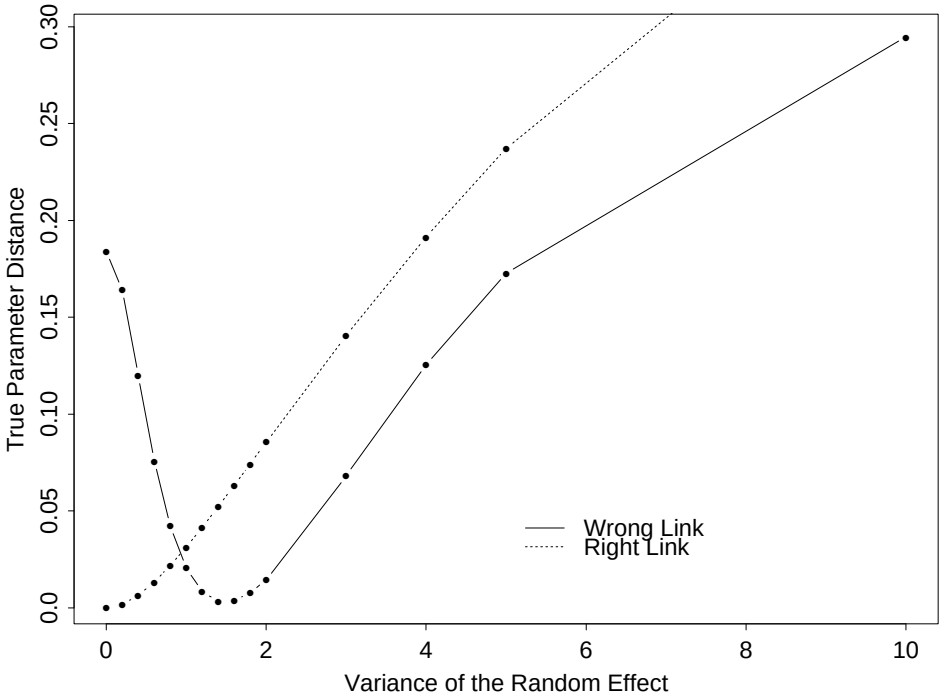

Figure 3.2: The distance between $\boldsymbol{\beta}_n^\star$ and $\boldsymbol{\beta}_0$ at the nonzero entries of $\boldsymbol{\beta}_0$ ($\boldsymbol{\beta}^*$) when the link function is correctly specified (dotted line) and when the link function is incorrectly specified (solid line)

from 0 when $\sigma_u^2 = 0$ to larger values when σ_u^2 gets large. When the link function is not correct, though, the distance $\|\boldsymbol{\beta}_n^\star - \boldsymbol{\beta}_0\|^2$ decreases at small σ_u^2 values and then increase when σ_u^2 continues to grow.

If we only compare the components of $\boldsymbol{\beta}_n^\star$ and $\boldsymbol{\beta}_0$ at the nonzero entries of $\boldsymbol{\beta}_0(\boldsymbol{\beta}^*,)$ there seems to be the same trend (Figure 3.2).

3.6 Conclusions

In this Chapter, we have presented the sufficient conditions under which the MLE $\hat{\boldsymbol{\beta}}_n$ for the working model converges in probability to a well defined limit $\boldsymbol{\beta}_n^\star$

when the model is misspecified and the number of parameters is going to infinity with the sample size, which have not been discussed in any existing paper in the literature. This limit $\boldsymbol{\beta}_n^\star$ may or may not be the true parameter, but is the parameter that minimizes the Kullback-Leibler distance between the true distribution of the data and the distribution of the data under the working model. Sufficient conditions for the MLE to be asymptotically normal according to Definition 1.12 were discussed. These results are elaborated in Section 3.3 and 3.4 under specific distributional assumptions on the random effect. When analytical approximations were not available, we numerically calculated quantities of our interest in a logistic-normal model and checked them with simulation studies. So far both the simulations and numerical calculations have supported our conjectures about the behavior of $\boldsymbol{\beta}_n^\star$ and the variance estimators of $\hat{\boldsymbol{\beta}}_n$.

Appendix A

Linear Algebra Results

Inequalities regarding to operator norm, Euclidean norm and trace of a matrix that are of particular use to us are discussed here. Most of the results are straightforward and can be derived directly from the definition.

Proposition A.1 *If the two $n \times n$ matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are both nonnegative definite, then*

$$tr[\mathbf{P}_1\mathbf{P}_2] > 0.$$

Proof: Since both $\mathbf{P}_1$ and $\mathbf{P}_2$ are nonnegative definite, their symmetric square roots, $\mathbf{P}_1^{1/2}$ and $\mathbf{P}_2^{1/2}$ exist and both are nonnegative definite. Therefore,

$$\mathrm{tr}[\mathbf{P}_1\mathbf{P}_2] = \mathrm{tr}[\mathbf{P}_1^{1/2}\mathbf{P_1}^{1/2}\mathbf{P}_2^{1/2}\mathbf{P}_2^{1/2}] = \mathrm{tr}[\mathbf{P}\mathbf{P}'] \geq 0$$

where $\mathbf{P} = \mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}$. □

A direct application of the proposition is

Corollary A.1 *If $\mathbf{P}_1 \leq \mathbf{P}_2$ and $\mathbf{P}_3$ is nonnegative definite matrix of the same dimension, then*

$$tr[\mathbf{P}_1\mathbf{P}_3] \leq tr[\mathbf{P}_2\mathbf{P}_3].$$

Proof: Since $\mathbf{P}_1 \leq \mathbf{P}_2$, the matrix $(\mathbf{P}_2 - \mathbf{P}_1)$ is nonnegative definite and by Proposition A.1,

$$\mathrm{tr}[\mathbf{P}_3\mathbf{P}_2 - \mathbf{P}_3\mathbf{P}_1] = \mathrm{tr}[\mathbf{P}_3(\mathbf{P}_2 - \mathbf{P}_1)] \geq 0.$$

Proposition A.2 *For $n\times 1$ vector $\mathbf{w}$, the norm of the rank-one matrix $\mathbf{ww}'$ satisfies*

$$\|\mathbf{ww}'\| \leq \|\mathbf{w}\|^2.$$

Proof: For any unit vector $\mathbf{v} \in \mathbf{R}^n$, by the Cauchy-Schwartz inequality, $\mathbf{v}'\mathbf{ww}'\mathbf{v} = (\mathbf{v}'\mathbf{w})^2 = \sum_{i=1}^n v_i^2 w_i^2 \leq \sum_{i=1}^n \mathbf{w}_i^2 = \|\mathbf{w}\|^2$. □

Proposition A.3 *For $n \times n$ symmetric nonnegative definite matrix $\mathbf{M}$,*

$$tr\mathbf{M} \leq n\lambda_{\max}(\mathbf{M}) = n\|\mathbf{M}\|.$$

Proof: The trace of a matrix is the sum of its eigenvalues, and for a nonnegative definite matrix all the eigenvalues are nonnegative, so $\mathrm{tr}\mathbf{M} \leq n\lambda_{\max}(\mathbf{M})$. The equality follows from Definition 1.7. □

Proposition A.4 *For full rank $n \times m$ matrix $\mathbf{X}$ and $n \times n$ diagonal matrix $\mathbf{D} = diag(d_1, \cdots, d_n)$,*

$$\|\mathbf{X}'\mathbf{DX}\| \leq \|\mathbf{D}\|\sqrt{\|\mathbf{X}'\mathbf{X}\|} \leq max_{1\leq k\leq n}|d_k|\sqrt{\|\mathbf{X}'\mathbf{X}\|}.$$

Proof: For any unit vector $\mathbf{v} \in \mathbf{R}^n$,

$$\|\mathbf{D}\| = \sup_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}'\mathbf{Dw}}{\|\mathbf{w}\|} \geq \frac{\mathbf{v}'\mathbf{X}'\mathbf{DXv}}{\|\mathbf{v}'\mathbf{X}'\|},$$

so

$$\mathbf{v}'\mathbf{X}'\mathbf{DXv} \leq \|\mathbf{D}\| \cdot \|\mathbf{v}'\mathbf{X}'\| = \|\mathbf{D}\|\sqrt{\mathbf{v}'\mathbf{X}'\mathbf{Xv}}$$

for any unit vector $\mathbf{v} \in \mathbf{R}^n$. Therefore

$$\|\mathbf{X}'\mathbf{DX}\| \leq \|\mathbf{D}\|\sqrt{\|\mathbf{X}'\mathbf{X}\|},$$

and $\mathbf{v}'\mathbf{D}\mathbf{v} = \sum_{k=1}^{n} v_k^2 d_k \leq max_{1\leq k\leq n}|d_k|$ since $\sum_{i=1}^{n} v_k^2 = 1$, so $\|\mathbf{D}\| \leq \max_k |d_k|$. □

Appendix B

Probability and Statistical Results

B.1 Sum of iid 0−Mean Sequence

For iid random variables ξ_i with $E[\xi_i] = 0$ and $E|\xi_i|^p < \infty$, one variant of the Burkholder Inequalities is

$$\left(E\left|\sum_{i=1}^{n}\xi_i\right|^p\right)^{1/p} \leq C_p\left(E\left[\sum_{i=1}^{n}\xi_i^2\right]^{p/2}\right)^{1/p}, \tag{B.1}$$

where C_p is a constant over n. The following proposition follows directly from (B.1):

Proposition B.1 *If iid random variables ξ_i satisfy $E[\xi_i] = 0$ and $E|\xi_i|^p < \infty$, then*

$$E\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right|^p \leq C_p^p\|\xi_1\|_p^p n^{-p/2} = O(n^{-p/2}). \tag{B.2}$$

Proof: By the triangle inequality, for iid 0-mean sequence ξ_i and $p > 2$,

$$\|\sum_{i=1}^{n}\xi_i^2\|_{p/2} \leq \sum_{i=1}^{n}\|\xi_i^2\|_{p/2},$$

which means that

$$\left(E\left[\sum_{i=1}^{n}\xi_i^2\right]^{p/2}\right)^{2/p} \leq \sum_{i=1}^{n}\left(E[\xi_i^2]^{p/2}\right)^{2/p} = n\|\xi_1\|_p^2.$$

Therefore, (B.1) becomes

$$\left(E\left|\sum_{i=1}^{n}\xi_i\right|^p\right)^{1/p} \leq C_p\left(E\left[\sum_{i=1}^{n}\xi_i^2\right]^{p/2}\right)^{1/p} \leq C_p\|\xi_1\|_p n^{1/2},$$

or $E\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right|^p \leq \frac{1}{n^p}C_p^p\|\xi_1\|_p^p n^{p/2} = C_p^p\|\xi_i\|_p^p n^{-p/2} = O(n^{-p/2})$. □

Theorem B.1 *Let* $\mathbf{M}$ *be a* $p_n \times p_n$ *matrix such that each element of* $\mathbf{M}$ *is the average of* n *iid* $0-$*mean random variables with finite* $(4r)^{th}$ *moment, i.e.*

$$\mathbf{M}_{kl} = \frac{1}{n}\sum_{i=1}^{n} \zeta_{kl}^{(i)},$$

where $\zeta_{kl}^{(i)}$ *are iid random variables with* $E\left[\zeta_{kl}^{(i)}\right] = 0$ *and* $E\left|\zeta_{kl}^{(i)}\right|^{4r} < \infty$ *for* $1 \leq k, l \leq p_n$. *If* $p_n = O(n^{\theta})$ *with* $0 < \theta < 1/4$ *and* $r > \theta/(1-4\theta)$, *then there exists* $\delta > 0$ *such that* $\|\mathbf{M}\| = O_p(p_n^{-1-\delta})$.

Proof: Let $\epsilon = r(1-4\theta) - \theta > 0$ and $0 \leq \delta \leq \epsilon/2r\theta$, then

$$\begin{aligned}
P\left[\|\mathbf{M}\| > p_n^{-1-\delta}\right] &\leq P\left[p_n \max_{k,l} |\mathbf{M}_{kl}| > p_n^{-1-\delta}\right] \\
&\leq P\left[\max_{k,l} |\mathbf{M}_{kl}| > p_n^{-2-\delta}\right] \\
&\leq p_n^2 \max_{k,l} P\left[|\mathbf{M}_{kl}|^{4r} > \left(\frac{1}{p_n^{2+\delta}}\right)^{4r}\right] \\
&\overset{Prop.B.1}{\leq} M_r p_n^2 n^{-2r} p_n^{8r+4r\delta} \\
&= O(n^{\theta(2+8r+4r\delta)-2r}) \to 0
\end{aligned}$$

where M_r is a constant that does not depend on k, l or n. □

B.2 Approximation of $\Phi(x)$ at large positive x

The cdf of standard normal, $\Phi(x)$, does not have closed form; at large positive x values, though, it can be approximated:

Proposition B.2 *For large, positive number* x,

$$1 - \Phi(x) - \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x} \sim -\frac{e^{-\frac{x^2}{2}}}{2x\sqrt{x^4 + x^2}}.$$

where $\Phi(\cdot)$ *is the standard normal cumulative distribution function.*

Proof: Let $\phi(\cdot)$ be the standard normal density function. For any $x > 0$,

$$\begin{aligned} 1 - \Phi(x) - \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x} &= 1 - \Phi(x) - \frac{\phi(x)}{x} \\ &= \int_x^\infty \phi(z)dz - \frac{1}{x}\int_x^\infty z\phi(z)dz \\ &= \int_x^\infty \phi(z)\left[1 - \frac{z}{x}\right]dz \\ &= -x\int_0^\infty \phi(x(w+1))wdw \\ &= -\frac{x}{\sqrt{2\pi}}\int_0^\infty \exp\{-\frac{x^2(1+w)^2}{2} + \ln w\}dw. \end{aligned} \tag{B.3}$$

Let

$$f(w) = -\frac{x^2(1+w)^2}{2} + \ln w.$$

Then

$$f'(w) = -x^2(1+w) + \frac{1}{w},$$

and

$$f''(w) = -x^2 - \frac{1}{w^2} < 0.$$

Let $w^* > 0$ be the unique point at which $f(w)$ is locally maximized. Then

$$f'(w^*) = -x^2(1+w^*) + \frac{1}{w^*} = 0,$$

and

$$w^* = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4}{x^2}}.$$

Then at $w = w^*$, since $f'(w^*) = 0$,

$$f(w) \approx f(w^*) + \frac{f''(w^*)(w - w^*)^2}{2},$$

When x is a large positive number, $4/x^2$ is small, so

$$w^* = -\frac{1}{2} + \frac{1}{2}(1 + \frac{4}{x^2})^{1/2} = -\frac{1}{2} + \frac{1}{2}(1 + \frac{1}{2}\frac{4}{x^2} + O(x^{-4})) = x^{-2} + O(x^{-4}) \approx x^{-2}.$$

Therefore,

$$\begin{aligned} f(w^*) &= \ln w^* - \frac{x^2(1+w^*)^2}{2} \\ &\approx \ln(x^{-2}) - \frac{x^2}{2}. \end{aligned} \tag{B.4}$$

$$f''(w^*) = -x^2 - (w^*)^{-2} = -x^2 - x^4(1+O(x^{-2}))^{-2} \sim -x^2 - x^4,$$

and

$$\Phi(w^*) \approx \frac{1}{2}.$$

Therefore (B.3) becomes

$$\begin{aligned} & -\frac{x}{\sqrt{2\pi}} \int_0^\infty e^{f(w)} dw \\ \sim & -\frac{x}{\sqrt{2\pi}} \int_0^\infty \exp\{f(w^*) + \frac{f''(w^*)(w-w^*)^2}{2}\} dw \\ = & -xe^{f(w^*)} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(w-w^*)^2}{2}(-f''(w^*)\} dw \\ = & -\frac{xe^{f(w^*)}}{\sqrt{-f''(w^*)}} \Phi(w^*) \\ \sim & -\frac{e^{-\frac{x^2}{2}}}{2x\sqrt{x^2+x^4}}. \end{aligned} \tag{B.5}$$

□

As a corollary of Proposition B.2, we get

$$1 - \Phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x}(1+O(x^{-2})).$$

BIBLIOGRAPHY

[1] Alan Agresti (2002), *Categorical Data Analysis* (2nd Edition). New York: Wiley Series in Probability and Statistics.

[2] Robert H. Berk (1972), *Consistency and Asymptotic Normality of MLE's for Exponential Models*. The Annals of Mathematical Statistics, Vol. 43, 193-204.

[3] N.E. Breslow (1984), *Extra-Poisson Variation in Log-linear Models*. Applied Statistics, Vol. 33, 38-55.

[4] N.E. Breslow (1990), *Tests of Hypotheses in Over-dispersed Poisson Regression and Other Quasi-likelihood Models*. Journal of the American Statistical Association, Vol.85, 565-571.

[5] D.L. Burkholder, *Distribution Function Inequalities for Martigales*. The Annals of Probability, Vol.1(1973), No.1, 19-42.

[6] D.R. Cox (1983), *Some Remarks on Overdispersion*. Biometrika, Vol.70, 269-274.

[7] C. Dean and J.F. Lawless (1989), *Tests for Detecting Over-dispersion in Poisson Regression Models*. Journal of the American Statistical Association, Vol. 84, 467-472.

[8] Christiana Drake and Allan McQuarrie (1995), *A Note on the Bias due to Omitted Confounders*. Biometrika, Vol. 82, 633-638.

[9] F. Eicker (1963), *Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions* The Annals of Mathematical Statistics, vol. 34, 447-456.

[10] Ludwig Fahrmeir and Heinz Kaufmann (1985), *Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models*. The Annals of Statistics, Vol. 13, 342-368.

[11] Shelby J. Haberman (1977), *Maximum Likelihood Estimates in Exponential Response Models*. The Annals of Statistics, Vol.5, 815-841.

[12] Xuming He and Qi-Man Shao (2000), *On Parameters of Increasing Dimensions*. Journal of Multivariate Analysis, Vol. 73, 120-135.

[13] P. J. Huber (1967), *The Behavior of Maximum Likelihood Estimators under Nonstandard Conditions*. Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability. Berkeley: University of California Press.

[14] S. Kullback and R.A. Leibler (1951), *On Information and Sufficiency*. Annals of Mathematical Statistics, Vol. 22, 79-86.

[15] T. L. Lai, Herbert Robbins and C. Z. Wei(1978),*Strong Consistency of Least Squares Estimates in Multiple Regression*. Procedings of National Academy of Science. Vol. 75, 3045-3046.

[16] John W. Lamperti (1996), *Probability: A survey of the Mathematical Theory* (2nd. Edition). New York: Wiley Series in Probability and Statistics.

[17] P.McCullagh and J.A. Nelder (1989), *Generalized Linear Models* (2nd. Edition). London: Chapman and Hall.

[18] Charles E. McCulloch and Shayle R. Searle (2000), *Generalized, Linear, and Mixed Models*. New York: Wiley.

[19] John M. Neuhaus (1998), *Estimation Efficiency with Omitted Covariates in Generalized Linear Models*. Journal of the American Statistical Association, Vol. 93, 1124-1129.

[20] Stephen Portnoy (1988), *Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity*. The Annals of Statistics, Vol.16, 356-366.

[21] C.R. Rao and Y. Wu (2001), *On Model Selection*. IMS Lecture Notes-Monograph Series, Vol.38, 1-64.

[22] C.R. Rao and Y. Wu (1989), *A strongly consistent procedure for model selection in a regression problem*. Biometrika, Vol. 76, 369-374.

[23] Shayle R. Searle (1971), *Linear Models* (Classic Edition). New York: Wiley.

[24] Jun Shao (1997), *An Asymptotic Theory for Linear Model Selection*. Statistica Sinica, Vol. 7, 221-264.

[25] Ritei Shibata (1981), *An Optimal Selection of Regression Variables*. Biometrika, Vol.68, 45-54.

[26] A. N. Shiryaev (1995), *Probability* (2nd Edition). New York Berlin Heidelberg: Springer-Verlag.

[27] Robert L. Strawderman and Anastasios A. Tsiatis (1996), *On Consistency in Parameter Spaces of Expanding Dimension: An Application of the Inverse Function Theorem*. Statistica Sinica, Vol.6, 917-923.

[28] A. W. van der Vaart (2000), *Aymptotic Statistics*(First Paperback Edition), Cambridge Series in Statistical and Probabilistic Mathematics.

[29] W.N. Venables and B.D. Ripley (2000), *Modern Applied Statistics with S-PLUS* (3rd Edition). New York: Springer-Verlag.

[30] A. H. Welsh (1989), *On M-processes and M-estimation*. Annals of Statistics, Vol. 17, 337-361.

[31] A. H. Welsh (1990), *Correction:"On M-processes and M-estimation"*. Annals of Statistics, Vol. 18, 1500.

[32] Halbert White (1982), *Maximum Likelihood Estimation of Misspecified Models*. Econometrica, Vol.50, 1-26.

[33] Jeffery R. Wilson (1989), *Chi-square Tests for Overdispersion with Multiparameter Estimates*. Applied Statistics, Vol. 38, 441-453.